Chapter 5: Foundations for inference

OpenIntro Statistics, 4th Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

Review

Last time: Point estimates and sampling variability (Ch. 5)

- Point estimates, sample size, variability
- Sampling distributions
- Binomial distribution
- Rule of succession for small samples

R Demo: Motivate CLT

Board work: Binomial distribution converges to Normal

R Demo: Binomial to Normal

Central Limit Theorem (Ch. 5.1-5.2)

Central Limit Theorem

Central limit theorem

Sample proportions will be nearly normally distributed with mean equal to the population proportion, p, and standard error equal to $\sqrt{\frac{p\ (1-p)}{n}}$.

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

- It wasn't a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population proportion.
- We won't go through a detailed proof of why $SE = \sqrt{\frac{p\ (1-p)}{n}}$, but note that as n increases SE decreases.
    - As n increases samples will yield more consistent $\hat{p}$s, i.e. variability among $\hat{p}$s will be lower.

CLT - conditions

Certain conditions must be met for the CLT to apply:

1. *Independence:* Sampled observations must be independent. This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and
   - if sampling without replacement, $n < 10\%$ of the population.

CLT - conditions

Certain conditions must be met for the CLT to apply:

1. *Independence:* Sampled observations must be independent. This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and
   - if sampling without replacement, $n < 10\%$ of the population.
2. *Sample size:* There should be at least 10 expected successes and 10 expected failures in the observed sample.
   This is difficult to verify if you don't know the population proportion (or can't assume a value for it). In those cases we look for the number of observed successes and failures to be at least 10.

When the conditions are not met...

- When either np or $n(1-p)$ is small, the distribution is more discrete.
- When np or $n(1-p) < 10$, the distribution is more skewed.
- The larger both np and $n(1-p)$, the more normal the distribution.
- When np and $n(1-p)$ are both very large, the discreteness of the distribution is hardly evident, and the distribution looks much more like a normal distribution.

Edfinity quiz

Extending the framework for other statistics

- The strategy of using a sample statistic to estimate a parameter is quite common, and it's a strategy that we can apply to other statistics besides a proportion.
    - Take a random sample of students at a college and ask them how many extracurricular activities they are involved in to estimate the average number of extra curricular activities all students in this college are interested in.
- The principles and general ideas are from this chapter apply to other parameters as well, even if the details change a little.