Ch. 19 - Statistical Inference

Def'n: Estimation is the assignment of value(s) to a population parameter based on a value of the corresponding sample statistic.

An estimator is a rule used to calculate an estimate.

An estimate is a specific value of an estimator.

Note: in this chapter, always assuming an SRS.

- Notation:
  - Let θ be a generic parameter.
  - Let $\hat{\theta}$ be an estimator – a statistic calculated from a random sample
  - Consequently, $\hat{\theta}$ is an r.v. with mean $E(\hat{\theta}) = \mu_{\hat{\theta}}$ and std. dev. $\sigma_{\hat{\theta}}$

Def'n: A point estimate is a *single number* that is our "best guess" for the parameter.

→ like a *statistic*, but more precise towards parameter estimation.

An interval estimate is an *interval of numbers* within which the parameter value is believed to fall.

Generic large sample confidence intervals:

Def'n: A confidence interval (CI) for a parameter θ is an interval estimate of plausible values for θ. With a chosen degree of confidence, the CI's construction is such that the value of θ is captured between the statistics L and U, the lower and upper endpoints of the interval, respectively.

The confidence level of a CI estimate is the success rate of the *method* used to construct the interval (as opposed to confidence in any particular interval). The generic notation is $100(1 - \alpha)\%$. Typical values are 90%, 95%, and 99%.

Ex19.1) Using 95% and the upcoming method to construct a CI, the method is "successful" 95% of the time. That is, if this method was used to generate an interval estimate over and over again with different samples, in the long run, 95% of the resulting intervals would capture the true value of θ.

Many large-sample CIs have the form:

$$\text{point estimate} \pm (\text{critical value}) \times (\text{standard error})$$

where "point estimate" is a statistic $\hat{\theta}$ used to estimate parameter θ,

"standard error" is a statistic $\hat{\sigma}_{\hat{\theta}}$ used to estimate std. dev. of estimator $\hat{\theta}$,

"critical value" is a fixed number z defined so that if Z has std. norm. dist'n, then $P(-z \leq Z \leq z) = 1 - \alpha =$ confidence level

The product of the "standard error" and "critical value" is the *margin of error*.

Note: critical value z often denoted by $z_{\alpha/2}$, where the notation reflects $P(Z > z) = \alpha/2$.

Ex19.2) if the confidence level is 95%, then $\alpha/2 = 0.025$ and $z_{0.025} = 1.96$.

(diagram drawn in class)

Table 19X0 – Critical values for usual confidence levels

100(1 – α)%	α	$\alpha/2$	$z_{\alpha/2}$
90%	0.10	0.050	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.58

The estimator $\hat{\theta}$ and its standard error $\hat{\sigma}_{\hat{\theta}}$ are defined so that, when the sample size n is sufficiently large, the sampling distribution of

$$\boxed{\frac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}} \doteq N(0,1)}$$

Thus,

$$P\left(-z \le \frac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}} \le z\right) \approx 1-\alpha$$

Algebraic manipulation yields

$$P\left(\hat{\theta} - z\hat{\sigma}_{\hat{\theta}} \le \theta \le \hat{\theta} + z\hat{\sigma}_{\hat{\theta}}\right) \approx 1-\alpha$$

Large Sample CI for Population Proportion

Recall the 3 rules regarding the general properties of the sampling distribution of $\hat{p}$.

Then, when n is large, a $(1 - \alpha)100\%$ CI for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Note that n being large also allows for the standard error to use $\hat{p}$ since p is unknown.

Assumptions:

1. $n\hat{p} \ge 15$ and $n(1-\hat{p}) \ge 15$,
2. the sample can be regarded as a random sample from the population of interest.

Ex19.3) A survey of 1356 random adults asked them to pick out the funniest city name in a list. 923 chose "Keokuk", 74 chose "Walla Walla", and 359 chose "Seattle". Let p be the proportion of all adults who would have answered "Seattle" had they been polled. Construct and interpret a 95% confidence interval for p.

$\boxed{\text{Assumptions: Random sample? Yes. } n\hat{p} = 359 \ge 15 \text{ and } n(1-\hat{p}) = 997 \ge 15\text{? Yes.}}$

- Parameter = p - Sample size: $n = 1356$
- Estimate: $\hat{p} = 359/1356 \approx 0.265$
- Standard Error = $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\dfrac{0.265(1-0.265)}{1356}} = 0.0120$
- Confidence Level is 95%, so critical value $z = 1.96$
- Interval is $0.265 \pm (1.96)(0.0120) = 0.265 \pm 0.023$ → (0.241, 0.288)

Direct interpretation: With 95% confidence, the true proportion of all adults who would have answered "Seattle" is between 0.241 and 0.288.

Never write $P(\hat{p}_L \le p \le \hat{p}_U) = 0.95$. Wrong conceptual interpretation.
Correct conceptual interpretation: If many samples were obtained and corresponding intervals calculated, about 95% of the intervals would cover p.

Note that the interval is not appropriate for small samples. Such an interval is obtainable, but not in this course.

The *margin of error* for a CI:
1. Increases as the confidence level increases.
2. Decreases as the sample size increases.

Ex19.4) Using the data from Ex19.3),

a) if confidence level is 99%,
$0.265 \pm (2.58)(0.0120) = 0.265 \pm 0.031$ → $(0.234, 0.296)$

b) suppose $n = 2712$, then

$$\text{std. error} = \sqrt{\frac{0.265(1-0.265)}{2712}} = 0.00847$$

$0.265 \pm (1.96)(0.00847) = 0.265 \pm 0.017$ → $(0.248, 0.282)$

Choosing the sample size:

Consider the CI as $\hat{p} \pm m$, where $m = z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

Recall that m is the margin of error. The width of the CI is $2m$. Now, we still want to see how large a sample size is required; hence, we rearrange to

$$n \approx \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{m}\right)^2$$

Round up n to next integer. Replace $\hat{p}$ by a prior estimate. If we don't have such information, then how to make n as large as possible? By choosing $\hat{p} = 0.5$, we maximize $\hat{p}(1-\hat{p})$ and get a conservative choice for n. This choice is most common. If, however, we expect $\hat{p}$ to be close to 0 or 1, say $\hat{p} \le 0.1$, then we could set $\hat{p} = 0.1$ to obtain a smaller n. In this situation, though, we would usually want a smaller m.

Ex19.5) If you wish to conduct a poll so that the margin of error is at most 3 percentage points with 99% confidence, what is the minimum sample size required?

If $\hat{p} = 0.5$, $\boxed{n \approx \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{m}\right)^2 = 0.5(1-0.5)\left(\frac{2.576}{0.03}\right)^2 \approx 1843.27 = 1844}$

Suppose we knew $\hat{p} \le 0.1$, then using $z = 2.576$, $n \approx 663.58 \approx 664$.