Chapter 7.2: Large-Sample Confidence Intervals for a Population Mean and Proportion

Instructor: Dr. Arnab Maity

Outline of this chapter:

- Large-Sample Interval for \( \mu \)
- Confidence Intervals for Population Proportion
Review of Chapter 7.1

1. Definition of a confidence interval

2. Interpretation of a confidence interval

3. A confidence interval for population mean \( \mu \) when
   - population distribution is normal
   - population standard deviation \( \sigma \) is known

4. Interval width and sample size calculation

Typically it is not reasonable to assume that the population standard deviation \( \sigma \) is known. Also, assuming that the population distribution is normal may not be realistic as well. In this chapter, we shall learn a method to construct confidence intervals where these two assumptions are not required.

**Assumption:** Sample size \( n \) is large.

We need to use the following result:

**The Central Limit Theorem (CLT):** Let \( X_1, X_2, \cdots, X_n \) be a random sample from a distribution with mean \( \mu \) and variance \( \sigma^2 \). Then if \( n \) is sufficiently large, \( \bar{X} \) has approximately a normal distribution with a mean of \( \mu \) and a variance of \( \sigma^2/n \).

- The parent distribution of \( X \)'s may not be normal (e.g., could be even discrete)
- The larger the value of \( n \), the better the approximation.

Also, we will use the following property:

**Result:** If \( n \) is sufficiently large, then the sample standard deviation \( S \) becomes “very close” to the population standard deviation \( \sigma \). Formally, we write

\[
S \xrightarrow{p} \sigma
\]

as \( n \) grows to infinity.
Large sample confidence interval for $\mu$

- Based on the two results (the CLT and the result about $S$), we can write that, if $n$ is sufficiently large, then we have

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

approximately has a $N(0,1)$ distribution.

- A large sample 100$(1-\alpha)%$ confidence interval for $\mu$ can be formed as

$$(\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}})$$

- Note that we replaced $\sigma$ by $S$ in the above formula

- This formula is valid regardless of the shape of the population distribution (for sufficiently large $n$). Text book suggest $n \geq 40$ as our rule of thumb.

- Interval width: $2z_{\alpha/2} \frac{S}{\sqrt{n}}$

1. Allowable mechanical properties for structural design of metallic aerospace vehicles requires an approved method for statistically analyzing empirical test data. The article Establishing Mechanical Property Allowables for Metals (J. of Testing and Evaluation, 1998: 293-299) used the data provided in Exercises 1.13 on tensile ultimate strength (ksi) as a basis for addressing the difficulties in developing such a method.

Use the accompanying descriptive statistics output to calculate a 98% confidence interval for the true average ultimate tensile strength.

$$n = 153, \quad \bar{x} = 135.39, \quad s = 4.59, \quad \text{Minimum} = 122.20, \quad \text{Maximum} = 147.70.$$
2. The alternating current (AC) breakdown voltage (the minimum voltage that makes an insulator react as a conductor) of an insulating liquid indicates its dielectric strength. One random sample of 48 observations on breakdown voltage (kV) gives

$$\sum_{i=1}^{48} x_i = 2626, \quad \sum_{i=1}^{48} x_i^2 = 144,950.$$ 

Construct a 95% confidence interval for the mean breakdown voltage.
Estimation of population proportion $p$

Often estimation of a population proportion is of interest. Typically, we observe the number of success $X$ among $n$ trials, where the probability of a trial being a success is $p$. Thus,

- Parameter: $p$
- Point estimator: $X/n$ (sample proportion of success)
- Point estimate: $x/n$ ($x$ = observed number of success, once a sample is taken)

**Example:** In 48 trials in a laboratory, 16 trials resulted in ignition of a particular type of substrate by a lighted cigarette. Let $p$ be the long-run proportion of all such trials that would result in ignition.

- Parameter:
  - $n =$
  - $x =$
- Point estimate:

**Question:** How to construct a confidence interval for $p$?

**Assumption:** The number of success $X$ out of $n$ trials is assumed to have a Binomial distribution $B(n, p)$

- Population mean: $E(X) =$
- Population standard deviation: $sd(X) =$

**Normal approximation to Binomial:** If $n$ (number of trials) is sufficiently large, then the sample proportion $\hat{P} = X/n$ approximate has a normal distribution with mean $p$ and variance $p(1 - p)/n$. In other words, if $n$ is large enough

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

- Mean of the estimated proportion $E(\hat{P}) =$
- Standard deviation $\sigma_P = sd(\hat{P}) =$

- Estimated standard deviation $S_p =$
A large-sample confidence Interval for a Population Proportion $p$

- If $n$ is sufficiently large, we can write

$$\frac{\hat{P} - p}{\sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}} \sim N(0, 1).$$

- A large sample 100$(1 - \alpha)$% confidence interval for $p$ can be formed as

$$\left(\hat{P} - z_{\alpha/2} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}, \quad \hat{P} + z_{\alpha/2} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}\right)$$

- What do we mean by “large $n$”: our rule of thumb is

$$n\hat{P} \geq 10 \quad \text{and} \quad n(1 - \hat{P}) \geq 10.$$ 

Pros and cons of the large sample interval

- Advantage: Simple to calculate, only need to have $\hat{P}$ and $n$

- Disadvantage: Performs poorly when the true proportion towards the boundary, that is, closer to 0 or 1. Also we need to check $n\hat{P} \geq 10$ and $n(1 - \hat{P}) \geq 10$.

Score interval for a Population Proportion $p$

- The 100$(1 - \alpha)$% score interval for $p$ can be formed as

$$\frac{\hat{P} + z_{\alpha/2}^2/(2n)}{1 + z_{\alpha/2}^2/n} \pm z_{\alpha/2} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n} + \frac{z_{\alpha/2}^2}{4n^2}} \frac{1 + z_{\alpha/2}^2/n}{1 + z_{\alpha/2}^2/n},$$

where the − corresponds to the lower confidence limit and the + to the upper confidence limit. (Textbook page 280, display (7.10))

- The score interval can be used with nearly all $n$ and $p$ values (no need to check large sample conditions!) and performs much better.
3. In 48 trials in a laboratory, 16 trials resulted in ignition of a particular type of substrate by a lighted cigarette. Let $p$ be the long-run proportion of all such trials that would result in ignition. Construct an approximate 95% confidence interval for $p$ using both the large sample and score intervals.
Minimum sample size necessary to ensure an interval width $w$ is

$$n \approx 4z_{\alpha/2}^2 \frac{p(1-p)}{w^2}.$$  

(a) Always round $n$ up
(b) Note that the above formula involves the unknown $p$
(c) Conservative approach: use $p = 0.5$ while calculating $n$, that is

$$n \approx 4z_{\alpha/2}^2 \frac{0.5^2}{w^2} = \frac{z_{\alpha/2}^2}{w^2}.$$

4. In the previous problem, what is the minimum sample size required to ensure a width of 0.10, and confidence level 0.95?