7.1 Basic Properties of Confidence Intervals

What’s Missing in a Point Estimate?

- Just a single estimate
- No idea how reliable this estimate is
- What we need:
  - how reliable it is
  - some measure of the variability

Ideas behind CI

- Realizing that a single estimate is insufficient
- Try to use an interval to cover our target
- Crucial aspects:
    - how reliable is this interval? - confidence level
    - how meaningful is this interval? - width

Introducing Confidence Intervals for population mean

- Start with the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- We have

$$E[\bar{X}] = \mu, \quad V[\bar{X}] = \frac{\sigma^2}{n}$$

- If X is normal,

$$P\left[\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right] = 1 - \alpha$$

Interval Derivation

$$\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}$$

$$-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Interpreting CI

- Notice only $\bar{X}$ is random, not μ
- Interval is random
- The probability that this random interval covers μ is $1-\alpha$
- If we repeat the experiment many many times, we should have $100(1-\alpha)\%$ of the intervals covering μ
- Typical $1-\alpha$ values: 90%, 95%, 99%...

Common Misconceptions

- Say 95%, $z_{0.025} = 1.96$, one interval (1.5,1.8) obtained from a sample
- False claims:
  - there is a 95% chance that μ is in this interval
  - 95% of the observations are in this interval
  - both are wrong!

Calculating the Bounds

- Determine the α needed
- Check the normal distribution requirement
- See if σ is available (most likely not, so approximation is needed)

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Examples

$$\bar{x} = 58.3, n = 50, \sigma = 3$$

- (a). 95%: $z_{0.025} = 1.96$, $z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{3}{\sqrt{50}} = 0.832$

$$(57.47, 59.13)$$

- (b). 99%: $z_{0.005} = 2.57,\ z_{0.005}\frac{\sigma}{\sqrt{n}} = 2.57 \times \frac{3}{\sqrt{50}} = 1.09$

$$(57.2, 59.39)$$

- more confidence, wider interval, less useful!

Role of Sample Size n

- Everything else being equal
  - n increased by 4 times - interval narrowed by half
  - n decreased by 4 times - interval enlarged by double
  - if a particular w (width) is required

$$n = \left(2z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2$$

7.2 Large-Sample CI for a Population Mean and Proportion

Conditions for Previous CI Discussion

- Focused on population mean μ
- Assume the population distribution to be normal
- Require random samples $(X_1, X_2, \ldots, X_n)$
- Population variance is assumed to be known σ^2
- Approximation of σ^2 by S^2 is often used, assuming large sample size n

Getting around the normal

- No reason to assume a normal distribution in practice (example, the proportion of successes)
- Often the estimator can be approximately normal **if the sample size is large enough** (CLT at work!)
- Formulas are similar, but the implications can be subtle

Large-Sample Interval for μ

- Estimator $\bar{X}$
- Approximately normal if n is large
- Standardize $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

- Large-Sample CI for μ

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Issues need to be resolved

- How large is large enough?
- Typical rule of thumb: $n > 40$
- Replacing σ with s
- Sample variance itself is random
- Choosing n to have a desired width? highly dependent on the sample variance

From μ to more general parameters

- Consider more general θ
- Do we still have an estimator that is approximately normal?
- Require $\hat{\theta}$
  - approximately normal
  - unbiased
  - an expression for $\sigma_{\hat{\theta}}$

Constructing Large-Sample CI

- If $\hat{\theta}$ satisfies these requirements, we have

$$P\left(-z_{\alpha/2} < \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) \approx 1-\alpha$$

- Still need to find an interval for θ !
  - does $\sigma_{\hat{\theta}}$ involve θ ?
  - or another parameter? available approximation?

CI for Population Proportion

- Parameter in question: population proportion

$$\theta = p$$

- Natural estimator: $\hat{p} = X/n$
- Expression for $\sigma_{\hat{\theta}}$

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

- Solve for the CI: $$p = \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + (z_{\alpha/2}^2)/n}$$

Two Different CI's for p

- More accurate

$$p = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + (z^2_{\alpha/2})/n}$$

- More simplified

$$p = \hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$$

- Advantages of the first: (1) closer to the the intended confidence level; (2) even works with sample sizes not so large!

One-Sided CI (Confidence Bounds)

- large-sample upper CB

$$\mu < \bar{x} + z_\alpha \frac{s}{\sqrt{n}}$$

- large-sample lower CB
-

$$\mu > \bar{x} - z_\alpha \frac{s}{\sqrt{n}}$$

7.3 Intervals Based on Normal Population Distribution

Key Assumptions

- Normal population distribution
- Population standard dev σ NOT known, so sample standard dev S is used
- Sample size not large!
- Problem: $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ is NOT normal!

A New Distribution

- What do we know about this distribution?
  - more variability than the normal - more spread
  - additional parameter: the number of degree of freedom (df) = n-1, where n is the sample size
- Student's t-distribution, or just the t-distribution

Properties of t Distribution

- Denote t_ν the density function curve for ν df
- Each t_ν curve is bell-shaped, centered at 0
- Each t_ν curve is more spread out than the standard normal curve
- As ν increases, the spread decreases
- As $\nu \to \infty$, the sequence of t_ν curves approaches the standard normal curve

Standardize the rv

- We use the following in place of Z

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

- T has the t distribution with df n-1 where n is the sample size
- As n becomes large, T can be replaced by Z

Probability Statement

- Replacing z_α with $t_{\alpha,\nu}$
- $t_{\alpha,\nu}$ is the location on measurement axis for which the area under the curve to the right of $t_{\alpha,\nu}$ is α
- $t_{\alpha,\nu}$ is called a t critical value
- Different df leads to different critical values

One-Sample t CI

- $100(1-\alpha)$ confidence interval for μ is

$$\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

- Upper confidence bound $\bar{x} + t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$
- lower confidence bound $\bar{x} - t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$
- Can we find the n to give a desired interval width?

Prediction Interval (PI)

- Already have a sample of size n ($X_1, X_2, \ldots, X_n$)
- About to pick up the next one (X_{n+1})
- What can we predict?
  - one shot: the sample mean
  - better prediction: an interval
- Prediction interval (PI) for a single observation

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s\sqrt{1 + \frac{1}{n}}$$

Tolerance Intervals

- Want to find an interval that captures certain percentage of the values in a normal distribution
- Allow some uncertain factor (confidence level)
- Tolerance interval for capturing at least k% of the values, with a confidence level 95%

$$\bar{x} \pm (\text{tolerance critical value}) \cdot s$$

- Critical values found in Appendix Table A.6

7.4 CIs for Variance and Standard Dev of a Normal Population

Estimating Variance

- σ always needed in estimation of other parameters
- A point estimate often suffices, but sometimes we would like to have a CI
- Distribution of S usually more complicated than that of $\bar{X}$

Normal Population Assumption

- We need to start from something familiar
- Population with normal distribution
- $\bar{X}$ is normal, but S is not
- chi-square χ^2 distribution
- Another parameter involved (df)

A Theorem

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$
- the rv

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2}$$

- has a chi-square distribution with n-1 df

chi-squared critical values

- similar ideas as $z_\alpha, t_{\alpha,\nu}$
- additional parameter - df (n-1 in our applications)
- $\chi^2_{\alpha,\nu}$ denotes the number of the measurement axis such that α of the area under the curve with ν df lies to the right of $\chi^2_{\alpha,\nu}$

What's different?

- χ^2 not symmetric, because it is positive, but unbounded in the positive direction
- Two ends:
  - lower - $\chi^2_{1-\alpha/2,n-1}$, upper - $\chi^2_{\alpha/2,n-1}$
- Interval for σ^2 , confidence level $100(1-\alpha)\%$

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

- Formulas for one-sided: $\alpha/2 \rightarrow \alpha$