Confidence Intervals

CS 3130 / ECE 3530: Probability and Statistics for Engineers

April 8, 2025

Motivation: Confidence Intervals

Point estimates give a single best guess of a parameter based on data. Confidence intervals estimate a range where the parameter likely falls, with a given probability.

Definition

Let $X_1, X_2, \ldots, X_n$ be a sample from distribution F with parameter θ.
A **100(1 - α)% confidence interval** for θ is a pair L_n, U_n such that:

$$P(L_n < \theta < U_n) = 1 - \alpha$$

Common choice: $\alpha = 0.05$ (95% CI).

CI for the Mean (Known Variance)

Assume $X_i \sim N(\mu, \sigma^2)$.

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Find critical value $z_{\alpha/2}$ such that:

$$P(-z_{\alpha/2} < Z_n < z_{\alpha/2}) = 1 - \alpha$$

Resulting CI:

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

Computing Critical Values in R

Use quantile of standard normal:

- `qnorm(1 - 0.5 * alpha)`
- For $\alpha = 0.05$: $z_{0.025} \approx 1.96$
- For $\alpha = 0.01$: $z_{0.005} \approx 2.58$

CI for the Mean (Known Variance)

According to the central limit theory, Even X_i is not Gaussian, we can still approximate

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Find critical value $z_{\alpha/2}$ such that:

$$P(-z_{\alpha/2} < Z_n < z_{\alpha/2}) = 1 - \alpha$$

Resulting CI:

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

Example: Wasatch Snowfall

- You want to estimate the average snowfall in the Wasatch front this year
- You take snowfall measurements at 40 different locations along the front.
- Experience from previous years indicates that there is a variance of 36 inches between measurements at these locations.
- You compute the average snowfall for the year is 620 inches.

What is a 95% confidence interval for the average snowfall?

Example: Wasatch Snowfall

Sample of $n = 40$, known variance $\sigma^2 = 36$, sample mean = 620 in.
95% CI:

$$620 \pm 1.96 \cdot \frac{6}{\sqrt{40}}$$

Margin of Error in Polls

- You may notice that when you see the results of a poll, there is often a statement such as "the margin of error for this poll is $\pm 3\%$".
- What does this mean, and how do they come up with this number?
- If we are asking people about a choice between two candidates, then we can model their answers as a *Bernoulli* distribution.

The goal of the poll is to estimate the parameter p, which is the *proportion* of people that will vote for candidate "1" over candidate "0".

Margin of Error in Polls

Model responses as Bernoulli(p).
Sample mean $\bar{X}_n \sim N(p, p(1-p)/n)$.
95% CI:

$$\bar{X}_n \pm 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}$$

Assume $p = 0.5$, $n = 100$:

$$\text{MOE} = 1.96 \cdot \frac{0.5}{10} = 0.098$$

Required Sample Size for MOE = 3%

Want:

$$1.96 \cdot \sqrt{\frac{0.25}{n}} = 0.03 \Rightarrow n \approx 1067$$

CI for the Mean (Unknown Variance)

So far the estimation of confidence intervals requires we know the true value of the *variance*. However, the true value of variance is typically *unknown* in practice.

CI for the Mean (Unknown Variance)

Let us replace the variance σ^2 with its unbiased estimator — sample variance $S_n = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$. Let us use:

$$T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

$$\bar{X}_n \pm t_{\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$$

CI for the Mean (Unknown Variance)

$$T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

$$\bar{X}_n \pm t_{\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$$

T_n is no longer $N(0,1)$. It is called Student's t-distribution with degree of freedom $n-1$.

The pdf for the t-distribution is very similar to the Normal (see the book, Wikipedia, or you can plot it in R). It is centered at zero and a symmetric "hill" shape, but it does have "heavier tails", meaning that it goes to zero slower than the Normal. As n gets large, the t-distribution converges in the limit to a standard normal $N(0, 1)$. The t-distribution has one parameter, the degrees of freedom m. The random variable T_n above has a t-distribution with degrees of freedom equal to $m = n - 1$. In terms of notation, this is written $T_n \sim t(n - 1)$.

Computing t Critical Values in R

try with online R platform for convenience:
`https://www.programiz.com/r/online-compiler/`

- `qt(1 - 0.5 * alpha, df = n - 1)`
- Example: `qt(0.975, df = 9)` returns 2.26

Example: Wasatch Snowfall with Unknown Variance

Sample mean = 620, $S_n^2 = 34$, $n = 40$.
Use $t_{\alpha/2, df=39}$ for 95% CI:

$$620 \pm t_{0.025,39} \cdot \frac{\sqrt{34}}{\sqrt{40}}$$

Problem:

Repeat the Wasatch snowfall analysis above, but this time you do not rely on previous estimates of the snowfall variance. You compute the variance in your measurements to be $S_n^2 = 34$ inches. How did the confidence interval change?