11. Confidence Intervals

Spring 2021

Matthew Blackwell

Gov 2002 (Harvard)

Interval estimation - what and why?

- $\widehat{\tau} = \overline{Y}_n - \overline{X}_n$ is our best guess about $\tau = \mu_y - \mu_x$
- But $\mathbb{P}(\widehat{\tau} = \tau) = 0$!
- Alternative: produce a range of plausible values instead of one number.
    - Hopefully will increase the chance that we've captured the truth.
- We can use the distribution of estimators to derive these intervals.

Definitions

- **Interval estimator** of θ is an interval between two statistics $C = [L, U]$.
  - $L = L(X_1, \dots, X_n)$ and $U = U(X_1, \dots, X_n)$ are functions of the data.
  - An estimator just like $\overline{X}_n$ but with two values.
  - Goal: to infer that C covers or contains the true value.
- **Coverage probability** of $C = [L, U]$ is the probability that C covers the true value θ.
- In math, $\mathbb{P}(L \leq \theta \leq U) = \mathbb{P}(\theta \in C)$
- Important: interval is the random quantity, not the parameter.

What is a confidence interval?

Definition

A $1 - \alpha$ **confidence interval** for a population parameter θ is an interval estimator $C = (L, U)$ with coverage probability $1 - \alpha$.

- The random interval (L, U) will contain the truth $1 - \alpha$ of the time.
- Ideally, we'd want a 100% confidence interval, but usually not possible.
- Extremely useful way to represent our uncertainty about our estimate.
  - Shows a range of plausible values given the data.

Simple confidence intervals

- Estimator $\hat{\theta}$ for θ with estimated standard error $\widehat{\mathsf{se}}[\hat{\theta}]$.
- Quick-and-dirty 95% confidence interval:

$$C = \left[\hat{\theta} - 2\widehat{\mathsf{se}}(\hat{\theta}),\ \hat{\theta} + 2\widehat{\mathsf{se}}(\hat{\theta})\right]$$

- More formal, normal-based $1 - \alpha$ confidence interval:

$$C = \left[\hat{\theta} - z_{1-\alpha/2}\widehat{\mathsf{se}}(\hat{\theta}),\ \hat{\theta} + z_{1-\alpha/2}\widehat{\mathsf{se}}(\hat{\theta})\right]$$

  - $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal.
- Student-t-based $1 - \alpha$ confidence interval:

$$C = \left[\hat{\theta} - q_{1-\alpha/2}\widehat{\mathsf{se}}(\hat{\theta}),\ \hat{\theta} + q_{1-\alpha/2}\widehat{\mathsf{se}}(\hat{\theta})\right]$$

  - $q_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the student t with df = r.

Deriving confidence intervals

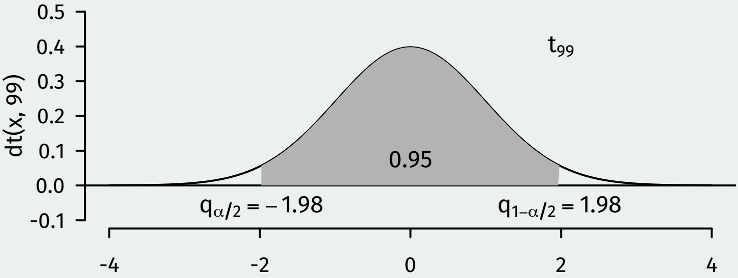
- How do we know the coverage of these confidence intervals?
- Sample mean: $\overline{X}_n$ with X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with $\widehat{\text{se}} = s/\sqrt{n}$.

$$T = \frac{\overline{X}_n - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

- T is a **pivotal quantity**: distribution doesn't depend on θ.
- If $q_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the t_{n-1}, we have

$$\mathbb{P}\left(-q_{1-\alpha/2} \leq \frac{\overline{X}_n - \mu}{s/\sqrt{n}} \leq q_{1-\alpha/2}\right) = 1 - \alpha$$

Finding the critical values



- How do we figure out what $q_{1-\alpha/2}$ will be?
- Intuitively, we want the q values that puts $\alpha/2$ in each of the tails.

$$\mathbb{P}(q_{\alpha/2} \leq T \leq q_{1-\alpha/2}) = 1 - \alpha$$

  - Because t is symmetric, we have $q_{\alpha/2} = -q_{1-\alpha/2}$
  - If $G(t)$ is the c.d.f. of T, then we have $q_{1-\alpha/2} = G^{-1}(1 - \alpha/2)$

Deriving the interval

- Let's work backwards to derive the confidence interval:

$$
\begin{aligned}
1-\alpha &= \mathbb{P}\left(-q_{1-\alpha/2} \le \frac{\overline{X}_n - \mu}{s/\sqrt{n}} \le q_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(-q_{1-\alpha/2} \times s/\sqrt{n} \le \overline{X}_n - \mu \le q_{1-\alpha/2} \times s/\sqrt{n}\right) \\
&= \mathbb{P}\left(-\overline{X}_n - q_{1-\alpha/2} \times s/\sqrt{n} \le -\mu \le -\overline{X}_n + q_{1-\alpha/2} \times s/\sqrt{n}\right) \\
&= \mathbb{P}\left(\overline{X}_n - q_{1-\alpha/2} \times s/\sqrt{n} \le \mu \le \overline{X}_n + q_{1-\alpha/2} \times s/\sqrt{n}\right)
\end{aligned}
$$

- Lower bound: $\overline{X}_n - q_{1-\alpha/2} s/\sqrt{n}$
- Upper bound: $\overline{X}_n + q_{1-\alpha/2} s/\sqrt{n}$
    - For 95% confidence interval with $n = 100$, $q_{1-\alpha/2} = 1.98$.
- Bounds are random! Not μ!

Asymptotic confidence intervals

- What about the $1-\alpha$ normal confidence interval:

$$C = \left[\hat{\theta} - z_{1-\alpha/2}\,\widehat{\text{se}}(\hat{\theta}),\ \hat{\theta} + z_{1-\alpha/2}\,\widehat{\text{se}}(\hat{\theta})\right]$$

- Asymptotically valid if our estimator is asymptotically normal so that:

$$\frac{\hat{\theta}_n - \theta}{\widehat{\text{se}}(\hat{\theta})} \xrightarrow{d} \mathcal{N}(0,1)$$

- Then as $n \to \infty$

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{\widehat{\text{se}}(\hat{\theta})} \leq z_{1-\alpha/2}\right) = \mathbb{P}(\theta \in C) \to 1-\alpha$$

- Again, $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ (`qnorm` in R)

CI for social pressure effect

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

```r
neigh_var <- var(social$voted[social$treatment == "Neighbors"])
neigh_n <- 38201
civic_var <- var(social$voted[social$treatment == "Civic Duty"])
civic_n <- 38218

se_diff <- sqrt(neigh_var/neigh_n + civic_var/civic_n)

## c(lower, upper)
c((0.378 - 0.315) - 1.96 * se_diff, (0.378 - 0.315) + 1.96 * se_diff)
```

```
## [1] 0.0563 0.0697
```

Interpreting the confidence interval

- **Caution:** a common **incorrect** interpretation of a confidence interval:
  - "I calculated a 95% confidence interval of [0.05,0.13], which means that there is a 95% chance that the true difference in means in is that interval."
  - This is WRONG.
- The true value of the population mean, μ, is **fixed**.
  - It is either in the interval or it isn't—there's no room for probability at all.
- The randomness is in the interval: $\overline{X}_n \pm 1.96s/\sqrt{n}$.
- Correct interpretation: **across 95% of random samples, the constructed confidence interval will contain the true value.**

Confidence interval simulation

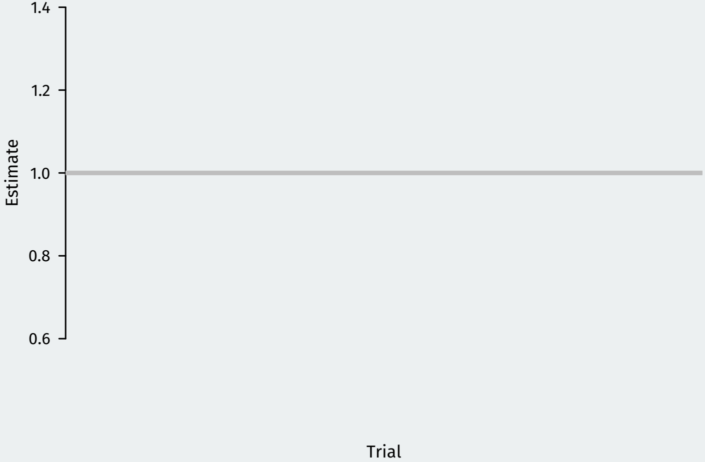
- Draw samples of size 500 (pretty big) from $\mathcal{N}(1, 10)$
- Calculate confidence intervals for the sample mean:

$$\overline{X}_n \pm 1.96 \times \widehat{\mathsf{se}}[\overline{X}_n] \rightsquigarrow \overline{X}_n \pm 1.96 \times s/\sqrt{n}$$

```r
sims<- 10000
cover <- rep(0, times = sims)
low.bound <- up.bound <- rep(NA, times = sims)
for(i in 1:sims){
    draws <- rnorm(500, mean = 1, sd = sqrt(10))
    low.bound[i] <- mean(draws) - sd(draws) / sqrt(500) * 1.96
    up.bound[i] <- mean(draws) + sd(draws) / sqrt(500) * 1.96
    if (low.bound[i] < 1 & up.bound[i] > 1) {
        cover[i] <- 1
    }
}
mean(cover)
```

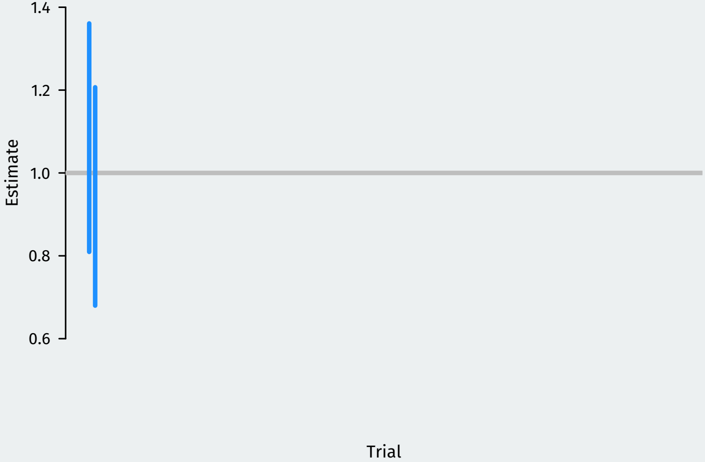
```
## [1] 0.95
```

Plotting the CIs



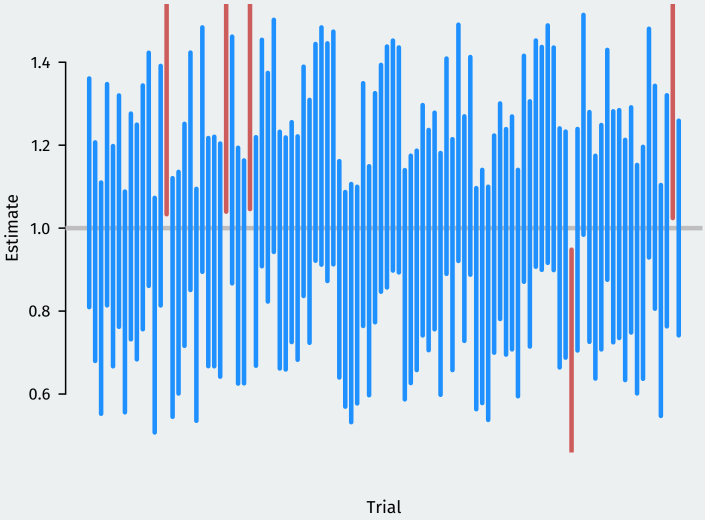
Plotting the CIs

Plotting the CIs



Plotting the CIs

Plotting the CIs



Question

- **Question** What happens to the size of the confidence interval when we increase our confidence, from say 95% to 99%? Do confidence intervals get wider or shorter?
- **Answer** Wider!
- Decreases $\alpha \rightsquigarrow$ increases $1 - \alpha/2 \rightsquigarrow$ increases $z_{\alpha/2}$

Inverting a hypothesis test

- 95% confidence interval: $\overline{X}_n \pm 1.96 \times s/\sqrt{n}$
- **CI/Test duality:** A $1-\alpha$ confidence interval contains all null hypotheses that we would not reject with a α-level test.
- Test of the null $H_0 : \mu = \mu_0$ at size α and reject when $|T| > z_{1-\alpha/2}$ where

$$T = \frac{\overline{X}_n - \mu_0}{s/\sqrt{n}}$$

- Reject when $\mu_0 > \overline{X}_n + z_{1-\alpha/2}\, s/\sqrt{n}$ or $\mu_0 < \overline{X}_n - z_{1-\alpha/2}\, s/\sqrt{n}$
  - $\rightsquigarrow$ reject any null outside the 95% confidence interval at size $\alpha = 0.05$
- CIs are a range of plausible values in the sense we cannot reject them as null hypotheses.