

Confidence intervals and testing

Day 2

Just Memorize These

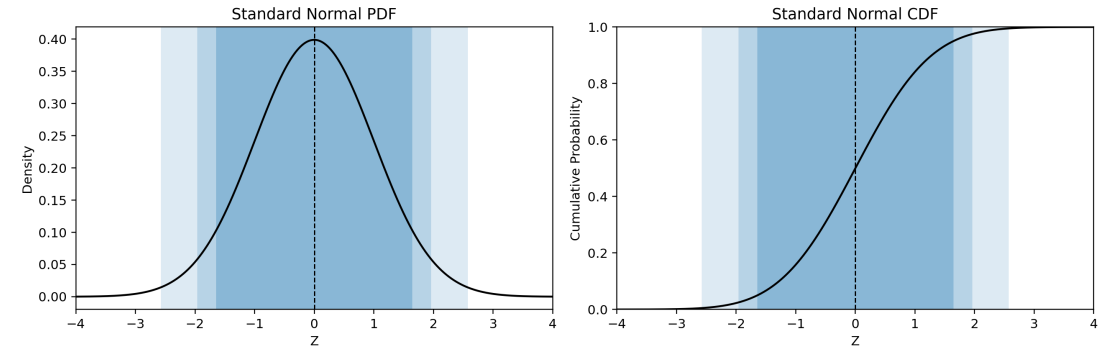
- Instead of taking the quantiles of our bootstrap sample, take the quantiles of the standard Normal, then invert the z-score normalization

Confidence	Lower	Upper	Level (α)
90%	-1.645	+1.645	.10
95%	-1.960	+1.960	.05
99%	-2.576	+2.576	.01

- So the 90% CI comes from this bound:

$$\left| \frac{\beta_0 - \hat{\beta}}{\widehat{SE}(\hat{\beta})} \right| \leq 1.645$$

$$\hat{\beta} - 1.645 \times \widehat{SE}(\hat{\beta}) \leq \beta_0 \leq \hat{\beta} + 1.645 \times \widehat{SE}(\hat{\beta})$$



What Is a Confidence Interval?

- Suppose our estimator is approximately normally distributed:

$$\frac{\hat{\beta} - \beta_0}{\widehat{SE}(\hat{\beta})} \sim N(0, 1)$$

- Choose a **critical value** c such that:

$$\Pr \left[\left| \frac{\hat{\beta} - \beta_0}{\widehat{SE}(\hat{\beta})} \right| \leq c \right] = 1 - \alpha$$

- What is random? The interval, because $\hat{\beta}$ changes with every sample. What is fixed? β_0 . **The 95% describes the procedure, not any one interval**
- Rearrange to get:

$$\Pr_{\hat{\beta}} \left[\hat{\beta} - c \times SE(\hat{\beta}) \leq \beta_0 \leq \hat{\beta} + c \times SE(\hat{\beta}) \right] = 1 - \alpha$$

- This is a statistical reliability claim, not a statement about the value of the true parameter β_0

Constructing Confidence Intervals

$$90\% : \left(\hat{\beta} - 1.645 \times \widehat{SE}(\hat{\beta}), \hat{\beta} + 1.645 \times \widehat{SE}(\hat{\beta}) \right)$$

$$95\% : \left(\hat{\beta} - 1.960 \times \widehat{SE}(\hat{\beta}), \hat{\beta} + 1.960 \times \widehat{SE}(\hat{\beta}) \right)$$

$$99\% : \left(\hat{\beta} - 2.576 \times \widehat{SE}(\hat{\beta}), \hat{\beta} + 2.576 \times \widehat{SE}(\hat{\beta}) \right)$$

Thought Experiment: Would the CLT work in each case?

For each one: does "estimate ± 1.96 SE" work? If not, what breaks down each time?

- 1. You survey 100 students that come in 10 friend groups whose members answer the same way
- 2. You average the profit ratio (this year's profit $\div$ last year's) across 100 stores, where many stores roughly broke even last year
- 3. A condition affects 1 in 100 people. You test 100 people and find 0 cases
- 4. You bootstrap the largest sale price in the Ames data
- 5. You poll 10,000 people, but only those who answer a landline at 2 pm

Confidence interval calculation exercise

- You run a regression and get a coefficient of 10 with a standard error of 5 for a key coefficient
- Compute 90/95/99% confidence intervals for the coefficient
- For which levels of confidence is a statistically significant difference from 0 detected?

CI to Hypothesis Test: TEST = INTERVAL

$$\Pr_{\hat{\beta}} \left[\hat{\beta} - c \times \text{SE}(\hat{\beta}) \leq \beta_0 \leq \hat{\beta} + c \times \text{SE}(\hat{\beta}) \right] = 1 - \alpha$$

- If we hypothesize β_0 takes some value, usually zero, this becomes a **hypothesis test**: is the hypothesized value β_0 in the interval or not?

$$\Pr_{\hat{\beta}} \left[\hat{\beta} - c \times \text{SE}(\hat{\beta}) \leq 0 \leq \hat{\beta} + c \times \text{SE}(\hat{\beta}) \right] = 1 - \alpha$$

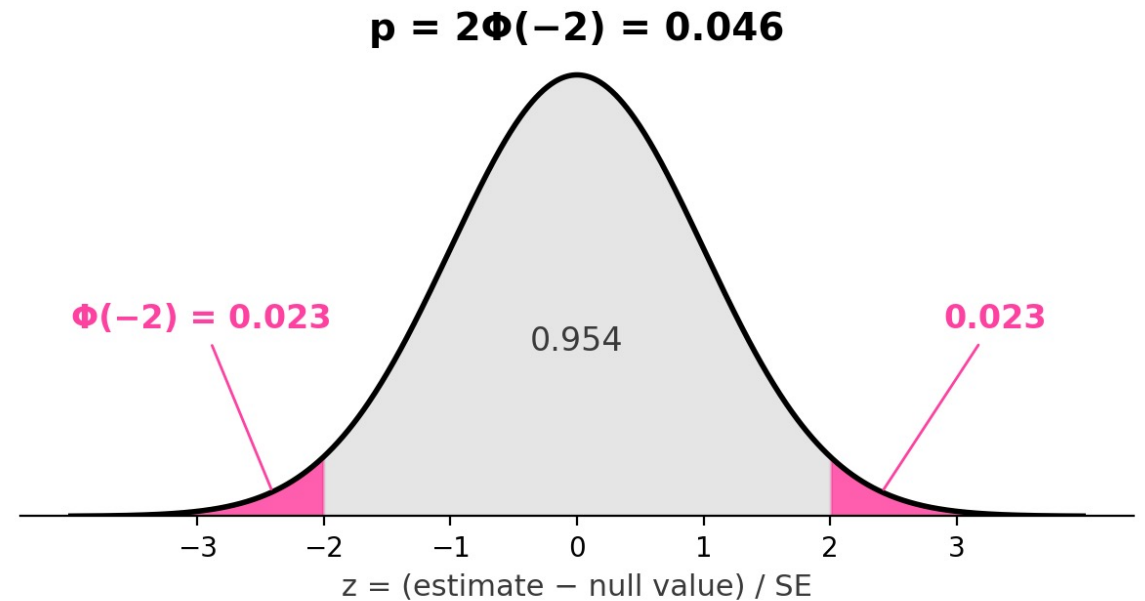
- Is β_0 (often 0) outside the CI? If so, reject H_0 ; if it is inside the CI, fail to reject H_0

From Test to p-Value

- The test on the last slide, like a confidence interval, gives a yes/no at one α . The **p-value** tells you every α at which you would reject
- Compute $z = (\text{estimate} - \theta_0) / \text{SE}$, then **$p = 2\Phi(-|z|)$** : the probability, **if the null were true**, of an estimate at least as far from θ_0 as ours
- Coefficient 10, SE 5: $z = 2$, so $p = 2\Phi(-2) = 0.046$. Reject at $\alpha = .10$ and $.05$, not at $.01$: the same answers the 90/95/99% CIs gave
- **Reject H_0 exactly when $p < \alpha$** , which is exactly when θ_0 is outside the $(1 - \alpha)$ CI. Same information, two ways to report it
- The p-value is **not** the probability that H_0 is true. It is computed assuming H_0 is true

Where Φ Comes In

- **Φ is the standard normal CDF**: $\Phi(z)$ is the area under the $N(0, 1)$ curve to the left of z
- By symmetry, $\Phi(-2) = 0.023$ is also the area to the right of $+2$. **$2\Phi(-|z|)$ adds both tails**: estimates at least this far from the null, in either direction
- It is the same function behind the critical values: $\Phi(-1.96) = 0.025$, so $2\Phi(-1.96) = 0.05$. That is where 1.96 comes from
- In Python: `from scipy.stats import norm`, then `p = 2 * norm.cdf(-abs(z))`
- Going the other way, Φ^{-1} gives critical values: `norm.ppf(0.975) = 1.96`



Remember, the area under the PDF curve to the left of z = CDF evaluated at z

The CI Is Not...

- the probability that β_0 is in the interval you computed
- It couldn't be this: we don't know the true β_0
- If β_0 was random, and had a distribution associated with it, then you'd use Bayes' Rule to observe data and update that distribution
- We often say things like, "Gravity is -9.81 m/s^2 ." We don't often say things like, "Gravity is normally distributed with mean -9.81 m/s^2 and standard deviation .0051"
- If you **did** talk like that, you'd be a Bayesian statistician, and we'll talk about Bayesian statistics at the end of the course

The CI Is...

- If we did this kind of analysis many times, what range of values is plausible for our estimates?
- A set of parameter values that our data do not allow us to reject, using a procedure designed to exclude the true parameter only a specified fraction of the time (the Type 1 error probability).
- The probability that the **random** CI covers the true β_0 is $1 - \alpha$; we **expect** the CI to cover the true β_0 with probability $1 - \alpha$
- If we repeatedly sampled data and constructed a new confidence interval each time, approximately 95% of those intervals would contain the fixed, unknown parameter β_0 (the bootstrapping connection is clear here)

Coverage You Can Count

Population {1, 3, 5}, draw $n = 2$ with replacement, so $\mu = 3$

Listing out all 9 possible samples and their $\bar{x}$

sample	(1,1)	(1,3)	(1,5)	(3,1)	(3,3)	(3,5)	(5,1)	(5,3)	(5,5)
$\bar{x}$	1	2	3	2	3	4	3	4	5

- 1. Our interval rule is $\bar{x} \pm 1.5$. Normally, we don't know exactly from how many samples would contain the true mean, because don't know what the true mean is. Instead, we use the CLT and normal curve to figure out the coverage of an interval. Here, we do know the true mean. **For how many of the 9 samples does the interval contain $\mu = 3$? What is the coverage?**
- 2. You drew (1, 1), so your interval is $[-0.5, 2.5]$. **What is the probability that this interval contains μ ?**

Comparing Two Groups: the SE of a Difference

- Drug trial: 25 patients on the standard drug (A) live an average of 10 months; 25 on a new drug (B) live 13. SD = 5 in each group, so each SE = $5/\sqrt{25} = 1$
- We care about the gap: $\bar{x}_B - \bar{x}_A = 3$ months. What is its SE?
- For **independent** groups, **variances add**: $V[\bar{x}_B - \bar{x}_A] = V[\bar{x}_B] + V[\bar{x}_A] = 1 + 1 = 2$, so **SE = $\sqrt{2} = 1.41$**
- **SEs do not add** ($1 + 1 = 2$ is wrong). In general: $SE(\text{gap}) = \sqrt{SE_A^2 + SE_B^2}$. Bootstrap version: resample each group separately, compute the gap each time, take the SD of the gaps

Comparing Two Groups: a CI for the Difference

- 95% CI for the gap: $3 \pm 1.96 \times 1.41 = 3 \pm 2.77 = (0.23, 5.77)$. 0 is outside, so the new drug adds months at $\alpha = .05$
- To get z, remember we will do estimated difference (3) – null difference (0) divided by SE:
- $z = 3 / 1.41 = 2.12$, and $p = 2\Phi(-2.12) = 0.03$
- But each drug's own 95% CI is (8.04, 11.96) for A and (11.04, 14.96) for B: **they overlap**
- **Overlapping CIs do not mean “no difference.”** But non-overlapping CIs DO mean a difference.
- Rule of thumb (independent groups, similar SEs): 95% CIs that just touch give $p \approx 0.006$, and they can overlap by about half an arm and still give $p < .05$

Paired Data: Take Differences First

- Crossover migraine trial: the **same 30 patients** spend a month on placebo (A) and a month on the drug (B). Averages: 10 and 11.5 migraine-free days; SE = 1 each
- As if the groups were independent: $SE(\text{gap}) = \sqrt{1 + 1} = 1.41$, $z = 1.5/1.41 = 1.06$, $p \approx 0.29$
- But patients with frequent migraines have few good days in both months, so each patient's two counts are correlated ($r \approx 0.8$)
- **Take each patient's difference $d = B - A$ first**, then do a one-sample CI on the 30 d's: $SE = s_d/\sqrt{30} = 0.63$, $CI = 1.5 \pm 1.24 = \mathbf{(0.26, 2.74)}$, $p \approx 0.02$
- Why: $V[\bar{x}_B - \bar{x}_A] = V[\bar{x}_A] + V[\bar{x}_B] - 2\text{Cov}(\bar{x}_A, \bar{x}_B)$. Positive correlation shrinks the SE, and two separate CIs can't show it

If you are getting confused about what to calculate...

you can't go wrong with the bootstrap version:

Two independent groups: resample each group separately, compute the gap each time, take the SD of the gaps.

Paired data (same units measured twice): compute each unit's difference, resample the differences, and take the SD of their means.

Inference routine

You should ALWAYS do

**Non-parametric exploration of
data **BEFORE** modeling**

You should ALWAYS do

Inference to evaluate estimators
NOT JUST prediction

Inference routine:

bootstrapping an estimate to understand its uncertainty

1. State the population parameter θ in words and define the estimator $\hat{\theta}$
2. Clean only the needed variables and report the resulting sample size(s).
3. Compute $\hat{\theta}$.
4. Generate at least $R=2,000$ nonparametric bootstrap estimates with a reproducible RNG. For two groups, resample independently within groups; for regressions, resample whole rows.
5. Estimate $\widehat{SE}(\hat{\theta})$ with the bootstrap standard deviation (`ddof=1`).
6. Construct centered-Normal 90%, 95%, and 99% CIs: $\hat{\theta} \pm z^* \widehat{SE}(\hat{\theta})$.
7. Test the stated null using $z = \frac{\hat{\theta} - \theta_0}{\widehat{SE}(\hat{\theta})}$ and $p = 2\Phi(-|z|)$. Decide at $\alpha = .10, .05, .01$.
8. Confirm that every test decision agrees with whether the matching CI contains θ_0 .
9. Interpret the estimate, interval, and test in context. These are observational data: do not turn association into causation.

1. **EXPLORE** the data non-parametrically before modeling
2. **ESTIMATION**: decide on what's important
3. **INFERENCE**: non-parametric bootstrap evaluation of parameters
4. **PREDICTION** (and quantifying the uncertainty of predictions)

Does It Pay for Itself?

- A retail chain pilots an energy-management system in 5 stores. It costs **\$200 per store per month**. Monthly savings on each store's electric bill: **\$100, \$300, \$400, \$500, \$700**
- 1. Compute $\bar{x}$, s^2 , and the estimated SE
- 2. Build a 95% CI with 1.96. Does the system save money at all? (Test H_0 : average savings = 0)
- 3. Does it pay for itself? Is the \$200 break-even inside or outside your CI?
- 4. The CFO writes: "There is a 95% chance the system saves between \$204 and \$596 per store." Rewrite it correctly
- 5. All 5 pilot stores are in Phoenix. What does your CI tell you about the stores in Maine?

Coding example

- Load the Metabric breast cancer dataset, and describe/plot `Overall Survival (Months)`.
- Describe a strategy for selecting variables that you'll use to predict `Overall Survival (Months)`, then implement your strategy and run your regression.
- Construct 90/95/99% non-parametric Bootstrap confidence intervals for the coefficients of your regression, as well as ones based on the Normal approximation using the standard deviation of the bootstrap sample as the standard error of the sampling distribution. How do they compare?
- Which variables are statistically significant, and at what levels?

Summary

Five Things to Think About

1. The parameter is fixed but unknown. We're trying to learn something about a property of the population.
2. Our estimates/confidence intervals are random variables of the data. Different samples produce different estimates, even for the same population doesn't change.
3. The standard error measures variability in our estimators, and bootstrapping gives us a practical way to estimate it.
4. The confidence interval translates variability into uncertainty about the parameter. Under the Normal approximation, that's $\pm 1.96 \times \text{SE}$. A hypothesized value outside the corresponding interval is rejected as implausible given the data and estimator.
5. The Bootstrap and the CLT are different approaches to understanding the underlying sampling distribution and making inferences about it, with their own strengths and weaknesses.

Summary

- The **standard error** is the standard deviation of the sampling distribution
- The **bootstrap** estimates the spread and shape of the sampling distribution from one sample
- The **CLT** says that, standardized, the sampling distribution is approximately $N(0, 1)$: that's why the bootstrap pictures all looked the same
- So we can replace bootstrap quantiles with $\pm 1.645, 1.96, 2.576$:

$$\hat{\theta} \pm c \times \widehat{SE}(\hat{\theta})$$

- A CI is a reliability claim about the procedure. A test asks whether the null value is inside the CI
- It breaks with dependence, heavy tails, small samples of skewed data, and estimates that aren't averages