CS109 Lecture 16: LLM Learning Guide

Bootstrapping

Summer 2026

Lecture 16: LLM Learning Guide

Before lecture: read the Lecture 16 slides (bootstrapping). Then open your favorite LLM and work through the concepts below **in order**. Each concept has a *Learn* prompt and a *Test me* prompt. Attempt the "Test me" questions yourself before asking the LLM to grade you.

How to get the most out of this:

- Tell the LLM your level: "I'm a student in Stanford's CS109 (probability for computer scientists). Explain at that level."
- After any answer, ask "why?" at least once and ask it to show each step.
- When it states a formula, ask it to define every symbol before continuing.
- Attempt practice yourself first, then say: "Here is my reasoning: _____. Where exactly am I wrong, if anywhere?"

Concept 1: Populations, samples, and sampling statistics

Learn: > "Explain the difference between a population distribution and a sample in statistics: we want facts about the population (its true mean μ and variance σ^2) but can only compute statistics from an i.i.d. sample of n draws. Explain why a statistic (like the sample mean) is itself a *random variable* with its own distribution, called the sampling distribution, and why that distribution is what error bars describe."

Test me: > "Give me several scenarios (polling, latency measurements, A/B tests) and make me identify: the population, the sample, the statistic being computed, and what quantity the error bars should describe. Check my answers."

Concept 2: The unbiased sample variance ($n-1$)

Learn: > "Explain why the naive sample variance $\frac{1}{n}\sum(x_i - \bar{X})^2$ systematically *underestimates* the true variance: the sample mean $\bar{X}$ is the value that minimizes the sum of squared deviations of the data, so deviations from $\bar{X}$ are smaller than deviations from the true μ. Show that dividing by $n-1$ instead of n (Bessel's correction) makes the estimator unbiased, and give the formula $S^2 = \frac{1}{n-1}\sum(x_i - \bar{X})^2$."

Test me: > "Give me a small dataset (5 or 6 numbers) and make me compute the sample mean and the unbiased sample variance by hand. Then ask me conceptual questions: which direction is the biased estimator wrong, why, and does the correction matter more for small or large n?"

Concept 3: Standard error of the mean

Learn: > "Explain the standard error of the mean: since $\mathrm{Var}(\bar{X}) = \sigma^2/n$, we estimate the standard deviation of the sample mean as $\sqrt{S^2/n}$. Explain the difference between the standard deviation of the *data* and the standard error of the *mean*, why the CLT makes $\bar{X}$ approximately normal, and how $\bar{X} \pm 2\,\mathrm{SE}$ gives an approximate 95% interval."

Test me: > "Give me a dataset summary (n, sample mean, sample variance) and make me compute the standard error and a 95% interval. Then ask me: what happens to the SE if n quadruples, and which of SE or sample standard deviation shrinks with more data?"

Concept 4: The bootstrap idea and algorithm

Learn: > "Explain the key insight of bootstrapping: the histogram of an i.i.d. sample is an estimate of the underlying distribution's PMF, so we can simulate 'new experiments' by resampling n values from our own sample *with replacement* (`np.random.choice(sample, n, replace=True)`). Walk through the full algorithm: repeat 10,000 times, recompute the statistic on each resample, and use the spread of the recomputed statistics as the sampling distribution. Explain why `replace=True` is essential and what breaks with `replace=False`."

Test me: > "Give me a tiny sample of 3 numbers and make me reason exactly about bootstrap resamples: how many equally likely ordered resamples there are, the probability the resample max equals the sample max, and the probability the resample mean equals the sample mean. Then make me explain in one sentence why sampling without replacement would make every resample's statistic identical."

Concept 5: Bootstrapping any statistic — and when it fails

Learn: > "Explain why the bootstrap works for essentially *any* statistic (median, variance, IQR, difference of means), unlike the CLT which only covers means and sums. Then explain its assumptions and failure modes: the samples must be i.i.d., and the sample histogram must be a decent estimate of the true distribution — which fails for long-tailed distributions where rare extreme values matter but are missing from the sample."

Test me: > "Quiz me with several statistics and datasets and ask, for each: would you use a CLT formula or the bootstrap, and is the bootstrap trustworthy here? Include at least one long-tailed example and one non-i.i.d. example, and check my reasoning."

Concept 6: The null hypothesis and p-values

Learn: > "Explain hypothesis testing with the bootstrap for a difference in means between groups A and B: state the null hypothesis (both groups come from the same distribution and any observed difference is sampling error), then describe the procedure — pool all the samples, repeatedly resample two groups of the original sizes from the pool, compute the difference in means each time, and report the p-value as the fraction of resampled differences at least as extreme as the observed one. Explain precisely what a p-value is and is not (it is *not* the probability the null is true)."

Test me: > "Give me an A/B test scenario with an observed difference of means and make me: state the null hypothesis, write pseudocode for the bootstrap p-value, and interpret p-values of 0.008 and 0.35. Grade my pseudocode line by line, especially which pool I resample from."

Wrap-up prompt (after all six concepts)

> "Give me one multi-part CS109-style problem in which I must (1) compute a sample mean, unbiased sample variance, and standard error from a small dataset, (2) write pseudocode to bootstrap the standard deviation of a statistic that has no closed-form formula, (3) reason exactly about resamples from a tiny sample, and (4) design a bootstrap hypothesis test for a difference between two groups and interpret the p-value. Let me solve every part, then grade each part, tell me which concept it tested, and tell me the single concept I should review most."

Note: bootstrapping was invented at Stanford by Bradley Efron in 1979, and it is how real experiments (A/B tests, scientific studies) put error bars and p-values on arbitrary statistics. It leans directly on last lecture's ideas: sampling, i.i.d. draws, and the CLT.

Bring any sticking points to lecture; the TAs and instructor will help you work through them on the worksheet.