Practice Quiz 4: Bootstrap and Permutation Tests

MSE 125 — Lectures 8–9

Use this practice quiz to prepare for Quiz 4 (Wednesday, April 29). The real quiz will have 2 questions in 10 minutes, closed-book. This practice set has 10 questions covering Lectures 8–9: the bootstrap procedure and percentile CI, the CLT and the normal approximation, when the normal approximation fails (medians, heavy tails, small n), sample-size planning, the permutation test and the p-value, the three p-value misconceptions, one-sided vs two-sided tests, and bootstrap vs permutation as tools.

Every concept tested on the real quiz appears somewhere on this practice set, with a different scenario.

Question 1. An analyst tries to build a bootstrap CI for the mean of a sample by resampling **without replacement** instead of with replacement.

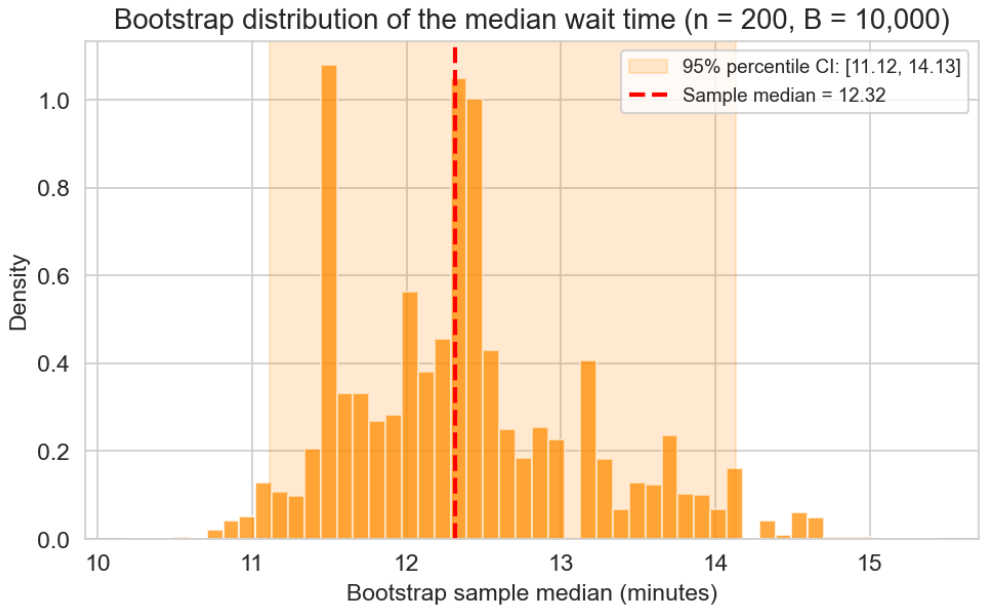
(a) Without computing anything, describe what the analyst's "bootstrap distribution" of the sample mean would look like across the 10,000 resamples. (1–2 sentences)

(b) Why does the bootstrap require sampling **with** replacement? Answer in 1 sentence.

Solution

(a) Every resample of size n drawn without replacement from a sample of size n is **just a permutation of the original** — every resample contains the exact same n values, in different orders. So every resample's mean is *identical* to the sample mean, the "bootstrap distribution" is a single spike with zero spread, and the implied SE is zero.
(b) With replacement is what produces variation across resamples — some observations appear twice, some don't appear at all, and that compositional jitter is exactly what mimics the variation we'd see across hypothetical re-runs of the study.

Question 2. A coffee chain wants a 95% confidence interval for the **median** wait time at a busy store. They collect $n = 200$ wait times and bootstrap the sample median over $B = 10{,}000$ resamples. The resulting distribution is below.



(a) Read off the 95% percentile CI for the median wait time: [_____, _____] minutes.

(b) Why does this bootstrap distribution look **lumpy** (many spikes), unlike the smooth bell curve we saw for bootstrapped means? Answer in 1 sentence.

(c) Could the chain have computed this CI using the normal approximation $\hat{\theta} \pm 1.96 \cdot \widehat{\mathrm{SE}}$ instead? Explain in 1 sentence.

Solution

(a) **[11.1, 14.1]** minutes (accept any reading within ±0.2 of the true endpoints 11.12 and 14.13).
(b) The sample median can only take values that actually appear in the resample — it lives on a *discrete grid* of order statistics — so the bootstrap distribution piles up on those grid points instead of spreading smoothly.
(c) No (or only awkwardly). The CLT gives a clean closed-form SE for the **mean**, but no equally simple formula exists for the median's SE — the asymptotic formula involves the unknown population density at the median, which itself would need to be estimated. The bootstrap is the practical tool here.

Question 3. A polling firm currently reports a margin of error of ±4 percentage points based on a sample of $n = 600$ voters.

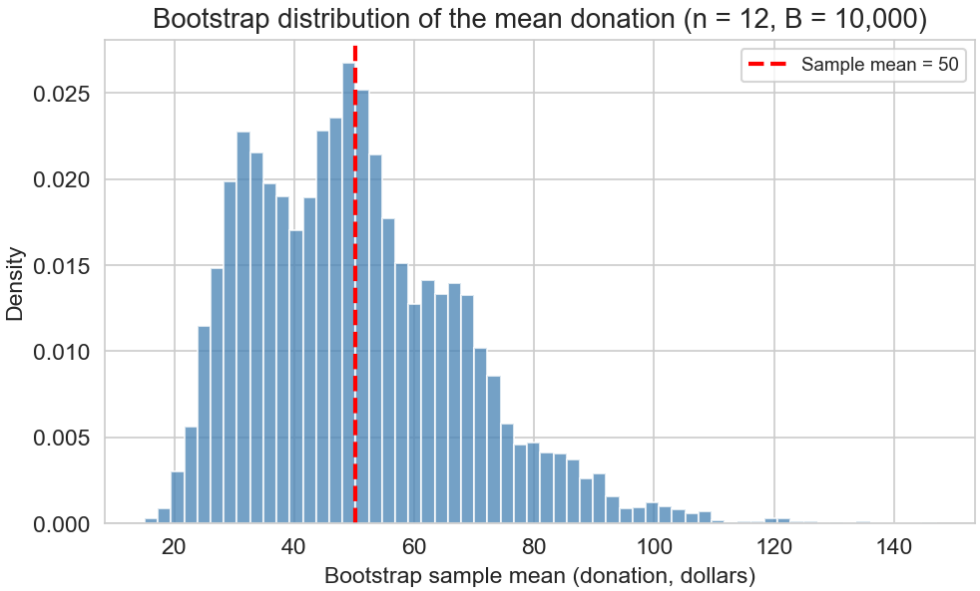
(a) The firm wants to halve the margin of error to ±2 percentage points. Approximately **how many voters** do they need to sample? _____

(b) In 1 sentence, justify your answer using the standard-error formula from the CLT.

Solution

(a) $n \approx$ **2,400** voters (a 4-fold increase, not 2-fold).
(b) The CLT says $\text{SE} = \sigma/\sqrt{n}$, so SE shrinks at rate $1/\sqrt{n}$ — to halve SE you need 4× the sample, not 2×.

Question 4. A small charity collects $n = 12$ donation amounts and bootstraps the sample **mean** over 10,000 resamples. The resulting distribution is below.



(a) The charity's analyst plans to report a 95% CI as $\bar{x} \pm 1.96 \cdot \widehat{\text{SE}}$ (the normal approximation). Looking at the bootstrap distribution above, what is the problem with this approach? (1–2 sentences)

(b) Should the analyst use the bootstrap percentile CI instead? Explain in 1 sentence — including any caveats.

(c) Suppose the charity ran this fundraising campaign in November. They want to use the CI to plan their **January** budget. Name **one** specific source of uncertainty about January donations that the bootstrap CI does **not** capture, and explain in 1 sentence why resampling cannot capture it.

Solution

(a) The bootstrap distribution is **visibly right-skewed** — the right tail is much longer than the left. The normal approximation $\bar{x} \pm 1.96 \cdot \widehat{\text{SE}}$ produces a **symmetric** interval, which mis-states the uncertainty in the same direction as the skew. The CLT hasn't kicked in yet at $n = 12$ on heavy-tailed donation data.
(b) Yes, the percentile CI is better here because it inherits the asymmetry of the bootstrap distribution, **but** the bootstrap isn't magic — at $n = 12$ the observed sample is itself a poor

picture of the population, so even the bootstrap CI should be treated as a signal of uncertainty rather than a final answer. Either collect more data or report results with explicit caveats.
(c) Examples (any one earns full credit):

- **Holiday-season effect** — November donations may include Giving Tuesday and end-of-year-tax-deduction motivations that don't apply in January.
- **A new competing nonprofit's campaign** launching in December.
- **A recession or stock-market move** between the campaigns.
- **Donor fatigue** — the same donors may give less in January after a big November ask.

In every case, the bootstrap **resamples rows of the November dataset**: it can only simulate sampling variation under November's actual conditions. It cannot simulate a world where donor behavior, the economy, or the competitive landscape has shifted, because no such draws appear in the data.

Question 5. A pharmaceutical company is designing a new clinical trial. They want to know **how many patients to enroll** so the resulting 95% CI for the mean change in cholesterol is no wider than ± 2 mg/dL. A junior analyst suggests: *"Let's just bootstrap our pilot data — that'll give us a CI, and we can read the sample size off the width."*

(a) Why can't the bootstrap answer the sample-size question? Answer in 1–2 sentences.

(b) Which tool does answer it, and what one input does it require that the bootstrap doesn't?

Solution

(a) The bootstrap can only resample *data we already have*. A sample-size calculation asks how many patients we'd need in a **future, larger** trial — there's no data yet at that size to resample. The bootstrap CI from the pilot tells us the SE at the pilot's n, not at any larger n.
(b) The **normal approximation** (CLT-based formula). It requires a *guess* at the population standard deviation σ (typically taken from pilot data or prior trials), which then plugs into $\text{SE} = \sigma/\sqrt{n}$ and lets us solve for the n that achieves the desired CI half-width $1.96 \cdot \sigma/\sqrt{n}$. This is the family of calculation NIH ran before ACTG 175 enrolled a single patient.

Question 6. A startup blog post reports: "Our bootstrap 95% CI for the conversion lift from the new checkout flow is **[+1.2%, +6.8%]**. So we are 95% confident that the true lift is between 1.2 and 6.8 percentage points." A statistician on the team replies: "That sentence is *ambiguous*. Depending on what you mean by '95% confident,' it's either correct or it's confusing two different things."

(a) Give a **charitable** reading under which the blog post's statement is correct, and a **strict** reading under which it is wrong. (2–3 sentences)

(b) Rewrite the sentence so the source of randomness is unambiguous.

Solution

(a) *Charitable (frequentist) reading:* "the procedure that produced this interval has 95% coverage across hypothetical repeats of the study, and this run produced [+1.2%, +6.8%]." Under this reading the probability is over the **sample** (which study you happened to run); the truth is a fixed-but-unknown number. The statement is correct.
Strict (Bayesian-sounding) reading: "given the data we collected, there is a 95% probability that the true lift sits between 1.2% and 6.8%." Under this reading the probability is over the **truth itself** (treated as a random variable). The bootstrap percentile CI does not deliver this — it uses no prior and runs no posterior calculation, so it cannot license a probability statement about the parameter. The textbook objection to the blog post is really objecting to *mixing* the two readings in one sentence.
(b) "Across many hypothetical repeats of this experiment, about 95% of the intervals built this way would contain the true conversion lift; for this run, the interval is [+1.2%, +6.8%]." (See Ch 8's "two epistemologies" callout for the alternative Bayesian phrasing, which would require a prior the bootstrap doesn't use.)

Question 7. A researcher reports $p = 0.01$ from a permutation test on a marketing experiment and says: "There's only a 1% chance the campaign **had no effect**."

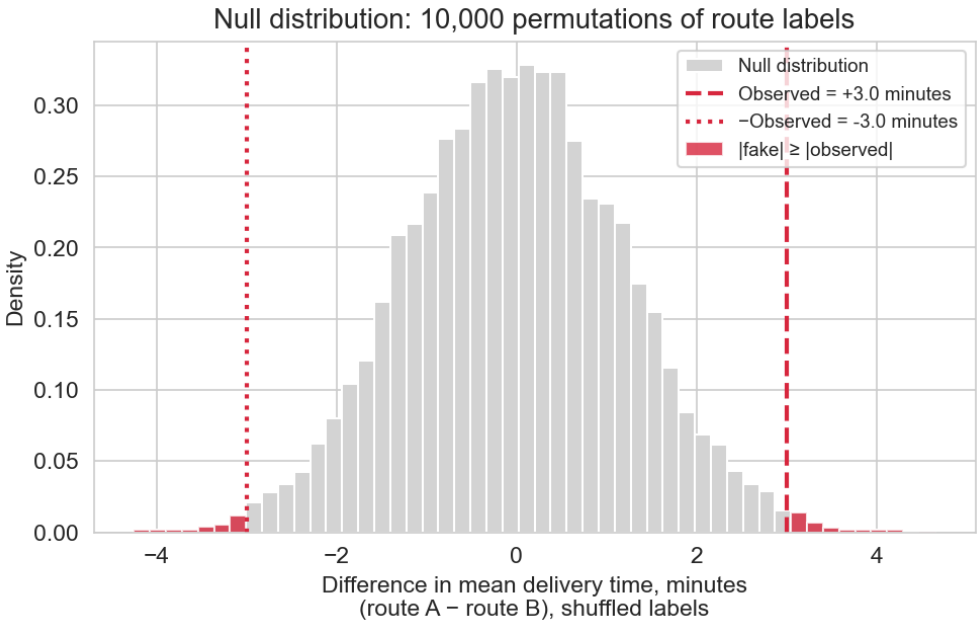
(a) The researcher's interpretation is wrong. State precisely what $p = 0.01$ does mean. (1 sentence)

(b) The researcher's statement and the correct interpretation are conditional probabilities running in opposite directions. Write both as $P(\cdot \mid \cdot)$ — using "extreme data" and "H_0 true" as your two events. Which one is the p-value, and which is the researcher's claim?

Solution

(a) $p = 0.01$ is the probability of seeing data **at least as extreme** as what was observed, **assuming the null hypothesis is true** (i.e., assuming the campaign had no effect). It says: *if* there were truly no effect, results this extreme would happen only 1% of the time.
(b) - The **p-value** is $P(\text{extreme data} \mid H_0 \text{ true})$. - The researcher's claim is $P(H_0 \text{ true} \mid \text{extreme data})$.
These are different conditionals. Going from one to the other requires Bayes' theorem and a prior probability on H_0 — neither of which the p-value uses. *(The "P(umbrella | raining) P(raining | umbrella)" intuition.)*

Question 8. A logistics company runs a randomized experiment to test whether a new delivery route is **faster** or **slower** than the current one (either direction matters operationally). They permute the route labels 10,000 times and build the null distribution of the difference in mean delivery times below.



(a) Roughly what fraction of permutations are at least as extreme as the observed +3.0 minute gap, in either direction? Circle one: ≈ 0.001 / ≈ 0.01 / ≈ 0.05 / ≈ 0.50

(b) After counting more carefully, the company finds that **119 of the 10,000 permutations** produced a fake effect at least as extreme as the observed +3.0 minutes (in either direction). Compute the **conservative** p-value using the $(n_{\text{extreme}} + 1)/(n_{\text{perms}} + 1)$ estimator: $p =$ _____

(c) At $\alpha = 0.05$, the two-sided test (circle one): **rejects** / **fails to reject** the null.

(d) The company also runs a bootstrap and gets a 99% CI for the mean difference of $[+1.4, +4.6]$ minutes. Without computing anything, what does this CI imply about the p-value at $\alpha = 0.01$? (1 sentence)

Solution

(a) $\approx$ **0.01** — the red-shaded tails together hold roughly 1% of the mass.
(b) $p = (119 + 1)/(10{,}000 + 1) = 120/10{,}001 \approx$ **0.0120**. The $+1/+1$ correction prevents reporting $p = 0$ from a finite simulation and keeps the estimator conservative.
(c) **Rejects** the null at = 0.05 (p 0.012 $<$ 0.05).
(d) The 99% CI excludes zero, so by **CI/test duality** the corresponding two-sided test rejects at = 0.01 — meaning the p-value is below 0.01. This is consistent with the $p \approx 0.012$ we computed in (b) (close to but slightly above 0.01; the small discrepancy is Monte Carlo noise between two separate simulations).

Question 9. A school district pre-registered a **two-sided** permutation test of whether a new math curriculum *changes* test scores compared to the previous curriculum. After running the test, the analyst sees scores went **up** in the treatment group with two-sided $p = 0.08$. The analyst writes: *"Since scores went up, only the upper tail matters — the one-sided p is 0.04, so we reject and*

recommend rolling out the new curriculum."

(a) Explain in 1–2 sentences what is wrong with the analyst's procedure.

(b) When an analyst uses the rule "look at the data, then declare one-sided in the direction of the observed effect, reject at $\alpha = 0.05$," what is the actual probability of falsely rejecting a true null? _____

Solution

(a) Post-hoc direction choice. The direction of a one-sided test must be chosen *before* seeing the data. The district pre-registered the two-sided test, which failed to reject. Switching to one-sided after seeing which way the effect went is equivalent to running two one-sided tests and reporting whichever one happens to be smaller — the analyst is "rescuing" significance from data that didn't earn it.
(b) ≈ 0.10 (or 2α, when the null distribution is symmetric — as it usually is with a difference-of-means test on roughly equal-sized groups). The procedure rejects whenever the observed statistic falls in *either* of the 5%-tails of the null, doubling the false-positive rate from 0.05 to roughly 0.10.

Question 10. Bootstrap or permutation? For each scenario, pick the right tool and answer the follow-up.

(a) An economist wants a 95% CI for the **median** household income in a sample of 800 households.

Tool: _____

Why this tool? (1 sentence)

(b) A health blog observes that people who drink 3+ cups of coffee daily have a 12% lower stroke rate. They run a permutation test on coffee-vs-no-coffee labels and get $p = 0.001$.

Tool used: permutation test. Does $p = 0.001$ license the conclusion *"coffee causes lower stroke rates"*? **Yes / No** — and why?

(c) A startup wants to test: "Is our new chatbot more accurate than the old one?" They run **the same 100 test queries** through each chatbot, scoring each response 1 (correct) or 0 (wrong). They plan a permutation test that **shuffles the chatbot labels across all 200 (chatbot, query) scores** and recomputes the difference in mean accuracy each time. State the assumption a permutation test requires under the null, and explain in 1–2 sentences whether it is satisfied here.

Solution

(a) Bootstrap. No clean closed-form SE exists for the median (the asymptotic formula involves the unknown population density at the median), so the bootstrap percentile CI is the practical tool. The CI quantifies the *precision* of the median estimate — sampling uncertainty over which 800 households happened to land in the sample.
(b) No. The permutation test is the right tool for testing whether the two groups' stroke rates differ, and $p = 0.001$ confirms they do — but **coffee drinking was not randomly**

assigned, so rejecting the null tells us only that the groups *differ*, not that coffee *causes* the lower stroke rate. Coffee drinkers may also exercise more, smoke less, see doctors more often — any of which could produce the gap. Random assignment is what licenses causal claims; without it, a small p-value is association, not causation.

(c) A permutation test requires **exchangeability under the null** — that the chatbot labels carry no information about the score, so any relabeling is equally plausible if the chatbots are equally accurate. **This is not satisfied here.** Each query has a fixed difficulty (some are easy, some are hard), so chatbot A's score on query 1 is *not* exchangeable with chatbot B's score on query 47 — shuffling labels across all 200 (chatbot, query) scores would scramble query difficulty into the comparison, inflating noise. The data are *paired* (same query $\times$ both chatbots), and the standard "shuffle all labels" recipe ignores the pairing. The right tool is a **paired** permutation test: for each query, randomly flip the A/B label, leaving the within-query pair intact.