STAT 140 Midterm 1 Final Review Guide

Coverage: Module 1-6 | Potential Outcomes, Assignment Mechanisms, Fisher, Neyman, Bayesian **Instructions**: Try each question first, then check the answer. Every question has a detailed explanation.

Part 1: 25 True/False Practice Questions

Q1

In the potential outcomes framework, $Y_i(1)$ and $Y_i(0)$ are fixed constants, not random variables.

Click to reveal answer

TRUE

In the Rubin Causal Model, potential outcomes are **fixed unknowns**, not random variables. For each unit i, $Y_i(1)$ and $Y_i(0)$ already "exist" before the experiment begins – we can only observe one of them. The only source of randomness is the **assignment vector W** (who gets assigned to treatment), not the potential outcomes themselves.

This is a core assumption of the Fisher/Neyman framework. In the Bayesian framework, missing potential outcomes are treated as random variables, but that is a modeling choice, not a statement about their intrinsic nature.

Q2

For a given unit i, the observed outcome Y_i^{obs} is a random variable.

Click to reveal answer

TRUE

Although $Y_i(0)$ and $Y_i(1)$ are fixed constants, $Y_i^{\text{obs}} = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0)$. Since W_i is random (determined by the randomization), Y_i^{obs} is also random – it "switches" between $Y_i(0)$ and $Y_i(1)$ depending on the assignment.

Q3

Under CRD, $P(W_i = 1) = N_t/N$ holds for all units i.

Click to reveal answer

TRUE

The symmetry of CRD guarantees that the **marginal probability** of each unit being assigned to treatment is the same, equal to N_t/N. This does not depend on any characteristic of the unit (potential outcomes, covariates, etc.), reflecting CRD's unconfoundedness.

Q4

Under CRD, the assignments of any two units $i \neq j$ are independent, i.e., $P(W_i = 1, W_j = 1) = P(W_i = 1) \cdot P(W_j = 1)$.

Click to reveal answer

FALSE

Assignments under CRD are **not independent**. Since the total number of treated units is fixed at N_t, knowing $W_i = 1$ reduces the remaining slots. Specifically:

$$P(W_j = 1 \mid W_i = 1) = \frac{N_t - 1}{N - 1} \neq \frac{N_t}{N} = P(W_j = 1)$$

Independence is a feature of **Bernoulli trials** (each unit flips an independent coin), not CRD. This is a high-frequency exam topic.

Q5

An advantage of Bernoulli randomization over CRD is that the number of treated units is fixed, preventing extreme imbalance.

Click to reveal answer

FALSE

It's the exact opposite. **CRD** fixes the number of treated units; **Bernoulli** does not. In a Bernoulli trial, each unit independently enters the treatment group with probability p, so in theory all units could end up in treatment (or all in control). CRD avoids this imbalance by fixing N_t.

Q6

The primary purpose of block randomization is to reduce covariate imbalance between groups, thereby improving estimation precision.

Click to reveal answer

TRUE

Block randomization first divides units into homogeneous "blocks" based on an important covariate (e.g., age, sex, baseline score), then independently performs CRD within each block. This ensures that treatment and control groups are perfectly balanced on that covariate within each block, reducing the estimation variance caused by covariate differences. Blocking is one of Fisher's three principles of experimental design.

Q7

In a paired randomized experiment with 20 units divided into 10 pairs, the number of possible assignment vectors is $\binom{20}{10}$.

Click to reveal answer

FALSE

In a paired experiment, one unit within each pair is independently and randomly assigned to treatment, so each pair has 2 choices. With 10 pairs, the total is $2^{10} = 1{,}024$ possible assignments. In contrast, $\binom{20}{10} = 184{,}756$ is the number of assignments under **CRD** (ignoring the pairing structure). Pairing **restricts** the randomization space, greatly reducing the number of possible assignments.

This distinction directly affects the size of the reference distribution when conducting a Fisher test.

Q8

When SUTVA is violated, we can still use $Y_i(0)$ and $Y_i(1)$ to define causal effects.

Click to reveal answer

FALSE

The first part of SUTVA (Stable Unit Treatment Value Assumption) is "no interference": unit i's potential outcomes depend only on i's own assignment. If SUTVA is violated (e.g., herd immunity in a vaccine trial), i's outcome also depends on other units' assignments. In that case, we need to write $Y_i(\mathbf{W})$ – a function of the entire assignment vector, not simply $Y_i(0)$ and $Y_i(1)$. The binary potential outcomes notation no longer applies.

Q9

ATE = 0 implies that every individual treatment effect $\tau_i = 0$.

Click to reveal answer

FALSE

ATE $= \frac{1}{N}\sum \tau_i = 0$ only requires the average of individual effects to be zero, not each one. For example, $\tau_1 = +10, \tau_2 = -10$ satisfies ATE = 0, but both individuals have non-zero effects. This is **treatment effect heterogeneity**, and it is the key distinction between the sharp null ($\tau_i = 0$ for all i) and the weak null (ATE = 0).

Q10

The sharp null hypothesis allows us to fill in the entire Science Table because it completely determines all missing potential outcomes.

Click to reveal answer

TRUE

The sharp null states $Y_i(1) = Y_i(0)$ for all i. This means each unit's outcome is the same regardless of assignment. So for treated units, $Y_i(0) = Y_i^{\text{obs}}$; for control units, $Y_i(1) = Y_i^{\text{obs}}$. Every missing value is determined, and the Science Table is completely filled. This is why the Fisher test can exactly enumerate all assignments.

Q11

The Fisher randomization test requires the assumption that potential outcomes follow a normal distribution.

Click to reveal answer

FALSE

The Fisher test **requires no distributional assumptions whatsoever**. It is "design-based" inference: the p-value comes entirely from the randomization mechanism, not from any assumption about the data distribution. This is why it is called an "exact test." In contrast, Neyman's confidence interval relies on the CLT (large-sample normal approximation).

Q12

The Fisher randomization test can only use $T = \bar{Y}_t - \bar{Y}_c$ as the test statistic.

Click to reveal answer

FALSE

The Fisher framework places **no restrictions** on the choice of test statistic. You can use the difference in means, Welch t-statistic, Wilcoxon rank sum, Kolmogorov-Smirnov statistic, or any function you define. Different statistics affect the **power** of the test (ability to detect real effects), but not its **validity** – the p-value is always exact.

Q13

A p-value of 0.03 means there is a 3% probability that the treatment has no effect.

Click to reveal answer

FALSE

This is one of the most common misinterpretations in statistics. The correct interpretation: **assuming the null hypothesis is true** (the treatment has no effect on anyone), the probability of observing a result as extreme as (or more extreme than) what we got is 3%.

It is not $P(H_0 \text{ is true} \mid \text{data})$, but rather $P(\text{data this extreme} \mid H_0 \text{ is true})$. Computing "the probability that the null is true" would require Bayes' theorem and a prior distribution.

Q14

Under CRD with $N = 4, N_t = 2$, the minimum possible p-value for a Fisher test is $1/6 \approx 0.167$, so it is impossible to reject any null hypothesis at the $\alpha = 0.05$ level.

Click to reveal answer

TRUE

$\binom{4}{2} = 6$ possible assignments, so the p-value can only take values $1/6, 2/6, 3/6, 4/6, 5/6, 1$. The smallest is $1/6 \approx 0.167 > 0.05$, so it is impossible to reach 0.05 significance. This illustrates a practical limitation of the Fisher test: **with too few units, the resolution of the p-value is too coarse**. You need more units to achieve smaller minimum p-values.

Q15

The Hodges-Lehmann estimator is the value of τ_0 that maximizes the Fisher p-value, and it equals the median of all pairwise differences.

Click to reveal answer

TRUE

These are two equivalent ways to compute the same thing: 1. **Test inversion approach**: Iterate over all τ_0, run a Fisher test $H_0 : \tau_i = \tau_0$ for each, and find the τ_0 with the largest p-value. 2. **Shortcut**: Compute each treated unit's outcome minus each control unit's outcome (all $N_t \times N_c$ differences), and take the median.

These are mathematically equivalent. Method 2 is an efficient computational shortcut for Method 1.

Q16

Neyman's difference-in-means estimator $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$ is unbiased for the ATE under CRD, even when treatment effects are heterogeneous.

Click to reveal answer

TRUE

The proof of unbiasedness relies only on CRD's symmetry: each unit has a N_t/N probability of entering the treatment group. This gives $E[\bar{Y}_t^{\text{obs}}] = \bar{Y}(1)$ and $E[\bar{Y}_c^{\text{obs}}] = \bar{Y}(0)$, so $E[\hat{\tau}] = \bar{Y}(1) - \bar{Y}(0) = \tau$.

This result **does not require** constant treatment effects. Heterogeneity affects the variance (through the $S^2(\tau)/N$ term), not unbiasedness.

Q17

Neyman's conservative variance estimator overestimates the true variance of $\hat{\tau}$ in expectation, but for any particular dataset, it may fall below the true variance.

Click to reveal answer

TRUE

"Conservative" is a property **in expectation**: $E[\widehat{V}] \geq \mathrm{Var}(\hat{\tau})$. But $\widehat{V}$ itself is a random variable (it varies with the assignment), so in any single experiment its realized value may happen to be lower than the true variance. Analogy: a thermometer that reads "high on average" doesn't mean every single reading is high – individual readings may happen to be low.

Q18

Neyman's 95% confidence interval means "there is a 95% probability that the true ATE falls within this interval."

Click to reveal answer

FALSE

The true ATE τ is a fixed constant, not a random variable, so there is no "probability" of it falling anywhere. The correct interpretation: **if we repeated this experiment many times and computed a 95% CI each time, approximately 95% of those intervals would contain the true τ**. The 95% describes the **reliability of the method**, not a property of any particular interval.

Q19

In Bayesian causal inference, single imputation underestimates uncertainty because it treats a guessed value as the true value.

Click to reveal answer

TRUE

Single imputation fills in each missing value with a single fixed number (e.g., the group mean), then pretends the Science Table is complete and computes the ATE. This **erases** our uncertainty about the missing values – in reality, A's outcome without coffee could be 60 or 75, but single imputation pretends it is definitely 67.5. The result is confidence intervals that are too narrow and standard errors that are too small. Multiple imputation corrects this by randomly sampling different fill-in values across many repetitions.

Q20

The Bayesian Bootstrap imputes missing values by drawing from a parametric normal distribution.

Click to reveal answer

FALSE

The Bayesian Bootstrap is **non-parametric**. It assumes no distributional shape. To impute $Y_i(0)$, it randomly draws from the **set of observed control outcomes**; to impute $Y_i(1)$, it draws from the **set of observed treatment outcomes**. This is the essence of “bootstrap” – using the existing data to pull itself up, without relying on any external distributional assumptions.

Q21

Fisher invented the Potential Outcomes Framework.

Click to reveal answer

FALSE

The history of the potential outcomes framework: **Neyman (1923)** first introduced the concept of potential outcomes in his doctoral dissertation (for agricultural experiments), but it was not widely disseminated. **Rubin (1974, 1978)** systematically developed and popularized the framework, making it the foundation of modern causal inference. It is therefore commonly called the **Rubin Causal Model** or the **Neyman-Rubin Framework**.

Fisher’s contributions were primarily in **randomization principles** and **randomization tests**, not the potential outcomes framework itself.

Q22

Causal inference is fundamentally a missing data problem.

Click to reveal answer

TRUE

This is a core insight of the Rubin Causal Model. In the Science Table, each unit has two potential outcomes $Y_i(0)$ and $Y_i(1)$, but we can only observe one – the other is permanently “missing.” If we could see the complete Science Table (all $2N$ potential outcomes), causal inference would be trivial: just compute $\tau_i = Y_i(1) - Y_i(0)$.

All methods of causal inference – Fisher’s test (using the sharp null to fill in missing values), Neyman’s estimation (bypassing missing values to directly estimate the mean difference), Bayesian imputation (modeling the missing values) – are fundamentally addressing this missing data problem.

Q23

Under CRD, all possible values of the assignment vector W form the set $\mathbb{W} = \{0, 1\}^N$, i.e., all 2^N possible N-dimensional binary vectors.

Click to reveal answer

FALSE

$\{0,1\}^N$ is the set of **all** N-dimensional binary vectors, containing 2^N elements. But under CRD, the number of treated units is fixed at N_t, so only vectors with exactly N_t ones have positive probability.

Correct description: - **Full set** $\mathbb{W} = \{0,1\}^N$ (all possible binary vectors, 2^N total) - **Positive-probability subset** $\mathbb{W}^+ = \{\mathbf{w} \in \{0,1\}^N : \sum_{i=1}^N w_i = N_t\}$ (vectors with exactly N_t ones, $\binom{N}{N_t}$ total)

Under CRD, $\mathbb{W}^+ \subset \mathbb{W}$, and each $\mathbf{w} \in \mathbb{W}^+$ has equal probability $1/\binom{N}{N_t}$.

This distinction does not exist under Bernoulli trials – there $\mathbb{W}^+ = \mathbb{W} = \{0,1\}^N$, and all 2^N vectors have positive probability.

Q24

Under CRD with $N = 10, N_t = 5$, the size of the assignment vector space $\mathbb{W}^+$ is $2^{10} = 1024$.

Click to reveal answer

FALSE

Under CRD, $\mathbb{W}^+$ contains only vectors with exactly $N_t = 5$ ones:

$$|\mathbb{W}^+| = \binom{10}{5} = 252$$

$2^{10} = 1024$ is the size of the assignment space under **Bernoulli trials**.

For this CRD: - Marginal probability of each unit being treated: $P(W_i = 1) = N_t/N = 5/10 = 0.5$ - Probability of each valid assignment: $P(\mathbf{W} = \mathbf{w}) = 1/252$ for $\mathbf{w} \in \mathbb{W}^+$

Q25

The ATT (Average Treatment Effect on the Treated) and ATE (Average Treatment Effect) always have the same expected value under CRD.

Click to reveal answer

FALSE

- ATE $= \frac{1}{N}\sum_{i=1}^N \tau_i$ (average effect across all units)
- ATT $= \frac{1}{N_t}\sum_{i:W_i=1} \tau_i$ (average effect among treated units only)

Under CRD, due to randomization, $E[\text{ATT}] = \text{ATE}$ – since each unit has an equal probability of entering the treatment group, the average effect in the treatment group equals the overall average effect in expectation.

However, for **any particular assignment**, ATT and ATE can differ, especially when treatment effects are heterogeneous. They are only equal in expectation.

The key word in this question is “always” – the realized values are not always the same, only the expected values are. If the question asked whether $E[\text{ATT}] = \text{ATE}$, the answer would be TRUE.

Part 2: 14 Core Concept Short Answers

Concept 1: What is the Potential Outcomes Framework?

Answer: Also called the Rubin Causal Model. The core idea: for each unit i, at any point in time there exist two **potential outcomes**:

- $Y_i(1)$: the outcome if i receives treatment
- $Y_i(0)$: the outcome if i does not receive treatment

The causal effect is defined as $\tau_i = Y_i(1) - Y_i(0)$. Due to the Fundamental Problem of Causal Inference, we can only observe one. All experimental design and inference methods aim to estimate causal effects as effectively as possible under this constraint.

Concept 2: What is SUTVA? Why is it important?

Answer: SUTVA (Stable Unit Treatment Value Assumption) has two parts:

1. **No Interference**: i’s outcome depends only on i’s own assignment W_i, not on others’ assignments.
2. **No Hidden Versions of Treatment**: there is only one version of treatment (no mixing of “high dose” and “low dose,” for example).

Why it matters: If SUTVA does not hold, the notation $Y_i(0)$ and $Y_i(1)$ itself becomes meaningless. For instance, in a vaccine trial with herd immunity – B’s health outcome depends on whether A also got vaccinated – we would need $Y_i(\mathbf{W})$, and complexity grows exponentially.

Concept 3: What are the key differences among CRD, Bernoulli Trial, and Block Randomization?

Answer:

Feature	CRD	Bernoulli	Block
Number treated	Fixed at N_t	Random (may be extremely unbalanced)	Fixed (within each block)
Assignment independence	Not independent (total constrained)	Independent	Independent across blocks, not within
Total assignments	$\binom{N}{N_t}$	2^N	$\prod 2^{n_k}$ (for pairs: 2^K)

Feature	CRD	Bernoulli	Block
Covariate balance	Balanced in expectation	Balanced in expectation but may deviate widely	Exactly balanced on blocking variable
Use case	General purpose	Simpler theoretical analysis	When important covariates are known

---

Concept 4: What is the Sharp Null? How does it differ from the Weak Null?

Answer:

- **Sharp Null** (Fisher): $H_0 : Y_i(1) = Y_i(0)$ for all i. Every individual's treatment effect is **exactly zero**.
- **Weak Null** (Neyman): $H_0 : \bar{Y}(1) - \bar{Y}(0) = 0$. The **average** treatment effect is zero, but individual effects can be heterogeneous.

Key distinction: The sharp null allows us to **fill in the entire Science Table** (because every missing value is determined), which is the prerequisite for the Fisher test to work. The weak null cannot do this, which is why Neyman's approach does not require a complete Science Table but instead makes inference on the sampling distribution of the estimator.

Concept 5: How do you interpret the p-value from a Fisher Randomization Test?

Answer: p-value = **assuming the sharp null is true**, the proportion of all equally-likely assignments that produce a test statistic at least as extreme as the one observed.

$$p = \frac{\text{number of assignments with } |T| \geq |T^{\text{obs}}|}{\text{total number of assignments}}$$

It is **not** "the probability that the null is true." A small p-value means: if treatment truly had no effect, the result we observed would be "too coincidental" – unlikely to arise from randomization alone.

Concept 6: What is Test Inversion? How does it construct a confidence interval?

Answer: Core idea: for each candidate treatment effect τ_0, run a Fisher test $H_0 : \tau_i = \tau_0$ for all i, and obtain a p-value $p(\tau_0)$.

$$\text{Confidence Interval} = \{\tau_0 : p(\tau_0) > \alpha\}$$

Collect all τ_0 values that "cannot be rejected" to form the $(1 - \alpha)$ confidence interval. It is **exact** and does not rely on the CLT. The downside is computational cost (each τ_0 requires a full Fisher test).

Concept 7: What is the Hodges-Lehmann estimator? How is it computed?

Answer: $\hat{\tau}_{HL}$ is the constant treatment effect **most compatible** with the data – i.e., the τ_0 that maximizes the Fisher p-value.

Shortcut computation: 1. For each treated unit i and each control unit j, compute the difference $Y_i^{\text{obs}} - Y_j^{\text{obs}}$ 2. There are $N_t \times N_c$ such differences 3. Take the **median** – that is $\hat{\tau}_{HL}$

The median is more **robust** than the mean and is not affected by outliers.

Concept 8: Why is Neyman's variance estimator called "conservative"?

Answer: The true variance has three terms:

$$\text{Var}(\hat{\tau}) = \frac{S^2(1)}{N_t} + \frac{S^2(0)}{N_c} - \frac{S^2(\tau)}{N}$$

The third term $S^2(\tau)/N$ is the variance of individual treatment effects, which is always ≥ 0. But we **cannot observe** it (since we never see individual effects), so Neyman simply drops it:

$$\hat{V}_{\text{Neyman}} = \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c}$$

Dropping a non-negative term means overestimating the variance → wider confidence intervals → more conservative (coverage $\geq 95\%$). When treatment effects are constant, $S^2(\tau) = 0$, and the conservatism vanishes.

Concept 9: What is the core idea behind the Bayesian approach to missing potential outcomes?

Answer: Fisher and Neyman treat missing values as **fixed unknown constants**. The Bayesian approach treats missing values as **random variables** and updates beliefs using Bayes' theorem:

$$p(\text{missing} \mid \text{observed}) \propto p(\text{observed} \mid \text{missing}) \times p(\text{missing})$$

The prior reflects pre-experiment beliefs, the likelihood measures how well hypothesized missing values fit the observed data, and the posterior is the final conclusion. The advantage is a **complete probability distribution**, not just a point estimate or interval.

Concept 10: When is each of the three inference approaches (Fisher, Neyman, Bayesian) most appropriate?

Answer:

Scenario	Best Method	Reason
Small sample ($N < 20$), want hypothesis test	**Fisher**	Exact, no large-sample approximation needed
Large sample, want point estimate and CI for ATE	**Neyman**	Simple formulas, CLT approximation is accurate
Reliable prior knowledge available	**Bayesian**	Can incorporate prior information
Need full distribution of treatment effect	**Bayesian**	Provides posterior distribution, not just point estimate
Want to test "no effect at all"	**Fisher**	Sharp null is most direct
Allow heterogeneous effects, only care about average	**Neyman**	Does not require constant effect assumption

---

Concept 11: Express Y_i^{obs} as a function of potential outcomes

Answer: The observed outcome can be expressed using the "Switching Equation":

$$Y_i^{\text{obs}} = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0)$$

Intuition: W_i acts like a switch. If $W_i = 1$ (treated), we see $Y_i(1)$; if $W_i = 0$ (control), we see $Y_i(0)$. This equation holds because of **SUTVA** – each unit's outcome depends only on its own assignment.

Equivalent notation: $Y_i^{\text{obs}} = Y_i(W_i)$

This equation connects three core concepts: potential outcomes $Y_i(0), Y_i(1)$, assignment W_i, and observed data Y_i^{obs}.

Concept 12: Describe the assignment vector sets $\mathbb{W}$, $\mathbb{W}^+$, and $P(W_i = 1)$

Answer: Using CRD with $N = 10, N_t = 5$ as an example:

Assignment Vector W $= (W_1, W_2, \dots, W_{10})$ is a 10-dimensional binary vector recording each unit's group assignment.

Full set $\mathbb{W}$: All possible 10-dimensional binary vectors, $|\mathbb{W}| = 2^{10} = 1024$. Includes extreme cases like "all treated" or "all control."

Positive-probability subset $\mathbb{W}^+$: Under CRD, only vectors with exactly 5 ones have positive probability. $|\mathbb{W}^+| = \binom{10}{5} = 252$.

Probability of each assignment: $P(\mathbf{W} = \mathbf{w}) = 1/252$ for $\mathbf{w} \in \mathbb{W}^+$, and 0 otherwise.

Marginal probability: $P(W_i = 1) = N_t/N = 5/10 = 0.5$ for all i.

Comparison of $\mathbb{W}^+$ across mechanisms:

Mechanism	$\mathbb{W}^+$	$\lvert\mathbb{W}^+\rvert$
CRD ($N = 10, N_t = 5$)	Vectors with exactly 5 ones	$\binom{10}{5} = 252$
Bernoulli ($p = 0.5$)	All binary vectors	$2^{10} = 1024$
Paired (5 pairs)	Vectors with exactly 1 one per pair	$2^5 = 32$

Concept 13: Why is causal inference a missing data problem?

Answer: In the Science Table, each unit has two potential outcomes $Y_i(0)$ and $Y_i(1)$, totaling $2N$ values. But we can only observe N of them (one per unit); the other N are permanently missing.

If the Science Table were complete (no missing values), causal inference would be simple arithmetic: $\tau_i = Y_i(1) - Y_i(0)$, ATE $= \frac{1}{N}\sum \tau_i$. **It is precisely because half the data is missing** that we need:

- **Fisher**: Use the sharp null hypothesis to fill in missing values (assume $Y_i(1) = Y_i(0)$)
- **Neyman**: Bypass missing values and directly exploit the properties of randomization to estimate ATE
- **Bayesian**: Build a probabilistic model for missing values and sample from the posterior distribution

Rubin's insight is that causal inference can be understood as a **special missing data problem**, where the missingness mechanism is the assignment mechanism, and randomization ensures the missingness is "random" (analogous to missing at random).

Concept 14: Describe SUTVA (Stable Unit Treatment Value Assumption)

Answer: SUTVA consists of two sub-assumptions:

(1) No Interference / No Spillover: Unit i's potential outcomes depend only on i's own treatment assignment W_i, not on other units' assignments:

$$Y_i(\mathbf{W}) = Y_i(W_i)$$

Violation example: Vaccine trial – if A gets vaccinated, A cannot transmit the disease to B, so B's health outcome depends on A's assignment, not just B's own.

(2) No Hidden Versions of Treatment: There is only one "version" of treatment. No mixing of "high dose" and "low dose."

Violation example: A study of "surgery vs. no surgery," but different surgeons have vastly different skill levels – then "surgery" actually has multiple versions, and $Y_i(1)$ is not uniquely determined.

Why SUTVA matters: If SUTVA does not hold, the simple notation $Y_i(0), Y_i(1)$ is insufficient – we need $Y_i(\mathbf{W})$, a function of the entire assignment vector. For N units, each unit would have 2^N possible potential outcomes (depending on everyone's assignment), and complexity explodes.

Part 3: Complete Worked Examples for Five Methods

> **Unified Scenario**: A researcher tests whether a **new learning App** improves math scores. 6 students participate; CRD assigns 3 to use the App (treatment) and 3 to use traditional methods (control).

Observed Data:

Student	W_i	Y_i^{obs} (Score)
A	1 (App)	85
B	1 (App)	78
C	1 (App)	92
D	0 (Traditional)	70
E	0 (Traditional)	74
F	0 (Traditional)	68

Summary statistics:

$$\bar{Y}_t = \frac{85+78+92}{3} = 85, \quad \bar{Y}_c = \frac{70+74+68}{3} = \frac{212}{3} \approx 70.67$$

$$T^{\text{obs}} = 85 - 70.67 = 14.33$$

Example 1: Fisher Randomization Test

Goal

Test the sharp null: the App has no effect on any student's score.

Step 1: State the hypothesis

$$H_0 : Y_i(1) = Y_i(0) \quad \text{for all } i = A, B, C, D, E, F$$

Step 2: Fill in the Science Table under sharp null

No treatment effect means each student's score is the same regardless of assignment:

Student	$Y_i(0)$	$Y_i(1)$
A	85	85
B	78	78
C	92	92
D	70	70
E	74	74
F	68	68

Step 3: Enumerate all $\binom{6}{3} = 20$ assignments

For each assignment, the treatment group takes the corresponding $Y(1)$ and the control group takes $Y(0)$ (which are identical under sharp null). Compute $T = \bar{Y}_t - \bar{Y}_c$:

#	Treated	Treated Values	$\bar{Y}_t$	Control Values	$\bar{Y}_c$	T
1	A,B,C	85,78,92	85.00	70,74,68	70.67	**14.33** (observed)
2	A,B,D	85,78,70	77.67	92,74,68	78.00	-0.33
3	A,B,E	85,78,74	79.00	92,70,68	76.67	2.33
4	A,B,F	85,78,68	77.00	92,70,74	78.67	-1.67
5	A,C,D	85,92,70	82.33	78,74,68	73.33	9.00
6	A,C,E	85,92,74	83.67	78,70,68	72.00	11.67
7	A,C,F	85,92,68	81.67	78,70,74	74.00	7.67
8	A,D,E	85,70,74	76.33	78,92,68	79.33	-3.00
9	A,D,F	85,70,68	74.33	78,92,74	81.33	-7.00
10	A,E,F	85,74,68	75.67	78,92,70	80.00	-4.33
11	B,C,D	78,92,70	80.00	85,74,68	75.67	4.33
12	B,C,E	78,92,74	81.33	85,70,68	74.33	7.00
13	B,C,F	78,92,68	79.33	85,70,74	76.33	3.00
14	B,D,E	78,70,74	74.00	85,92,68	81.67	-7.67
15	B,D,F	78,70,68	72.00	85,92,74	83.67	-11.67
16	B,E,F	78,74,68	73.33	85,92,70	82.33	-9.00
17	C,D,E	92,70,74	78.67	85,78,68	77.00	1.67
18	C,D,F	92,70,68	76.67	85,78,74	79.00	-2.33
19	C,E,F	92,74,68	78.00	85,78,70	77.67	0.33
20	D,E,F	70,74,68	70.67	85,78,92	85.00	**-14.33**

Step 4: Compute p-value

One-sided ($T \geq 14.33$): Only #1 → $p = 1/20 = 0.05$

Two-sided ($|T| \geq 14.33$): #1 and #20 → $p = 2/20 = 0.10$

Step 5: Conclusion

One-sided p = 0.05, exactly on the boundary. Two-sided p = 0.10 > 0.05, cannot reject.

Interpretation: The observed difference of 14.33 points is the most extreme among all 20 assignments, but because the sample size is so small ($\binom{6}{3} = 20$), the Fisher test has limited resolution. The result suggests the App may be effective, but the evidence is not strong enough.

Example 2: Hodges-Lehmann Estimator

Goal

Find the constant treatment effect value most compatible with the data.

Step 1: Compute all pairwise differences

Each treated unit's score minus each control unit's score (3 × 3 = 9 differences):

	D (70)	E (74)	F (68)
A (85)	85-70 = **15**	85-74 = **11**	85-68 = **17**
B (78)	78-70 = **8**	78-74 = **4**	78-68 = **10**
C (92)	92-70 = **22**	92-74 = **18**	92-68 = **24**

Step 2: Sort all 9 differences

$$4, 8, 10, 11, 15, 17, 18, 22, 24$$

Step 3: Take the median

9 values; the median is the 5th = **15**

$$\hat{\tau}_{HL} = 15$$

Interpretation

Our best estimate of the App's effect is: **using the App improves scores by an average of 15 points**.

Note that $\hat{\tau}_{HL} = 15$ differs slightly from $T^{\text{obs}} = 14.33$. T^{obs} is the difference in means; $\hat{\tau}_{HL}$ is the median of pairwise differences. The median is more robust and not affected by extreme values.

Verification: Relationship to Test Inversion

If we ran a Fisher test for each τ_0 and plotted the $p(\tau_0)$ curve, the p-value is largest at $\tau_0 = 15$ – it is the effect value that is "least likely to be rejected."

Example 3: Neymanian Inference

Goal

Estimate the ATE and construct a confidence interval, **without assuming** constant treatment effects.

Step 1: Point estimate

$$\hat{\tau} = \bar{Y}_t - \bar{Y}_c = 85 - 70.67 = 14.33$$

Step 2: Compute sample variances

Treatment group: {85, 78, 92}, $\bar{Y}_t = 85$

$$s_t^2 = \frac{(85-85)^2 + (78-85)^2 + (92-85)^2}{3-1} = \frac{0+49+49}{2} = 49$$

Control group: {70, 74, 68}, $\bar{Y}_c = 70.67$

$$s_c^2 = \frac{(70-70.67)^2 + (74-70.67)^2 + (68-70.67)^2}{3-1} = \frac{0.449+11.089+7.129}{2} = \frac{18.667}{2} = 9.333$$

Step 3: Conservative variance estimate

$$\hat{V}_{\text{Neyman}} = \frac{s_t^2}{N_t} + \frac{s_c^2}{N_c} = \frac{49}{3} + \frac{9.333}{3} = 16.333 + 3.111 = 19.444$$

Step 4: Standard error

$$SE = \sqrt{19.444} = 4.410$$

Step 5: 95% confidence interval

$$CI = \hat{\tau} \pm 1.96 \times SE = 14.33 \pm 1.96 \times 4.410 = 14.33 \pm 8.644$$

$$CI = [5.69,\ 22.97]$$

Interpretation

- **Point estimate**: The App improves scores by an average of 14.33 points
- **95% CI**: $[5.69, 22.97]$, does not contain 0 → at the 95% level, the App's effect is significant
- The effect is roughly between 6 and 23 points; the wide interval reflects the uncertainty from the small sample size (only 6 students)

Note

Neyman CI relies on the CLT, but here $N = 6$ is very small, so the normal approximation may be inaccurate. For such small samples, Fisher's exact CI (via test inversion) is more reliable. This example is primarily for demonstrating the calculation workflow.

Example 4: Bayesian Inference (Parametric Multiple Imputation)

Goal

Obtain a posterior distribution of the ATE by modeling the missing potential outcomes.

Step 1: Identify missing values

Student	$Y_i(1)$	$Y_i(0)$	What's missing
A	85	?	$Y_A(0)$
B	78	?	$Y_B(0)$
C	92	?	$Y_C(0)$
D	?	70	$Y_D(1)$
E	?	74	$Y_E(1)$
F	?	68	$Y_F(1)$

Step 2: Specify the model (prior + likelihood)

Assume:

- $Y_i(0) \sim N(\mu_0, \sigma_0^2)$
- $Y_i(1) \sim N(\mu_1, \sigma_1^2)$

Estimate parameters from observed data:

- $\hat{\mu}_0 = 70.67, \quad \hat{\sigma}_0^2 = 9.333$
- $\hat{\mu}_1 = 85.00, \quad \hat{\sigma}_1^2 = 49.0$

Step 3: Multiple imputation (5 rounds illustrated)

Each round, randomly draw a value for each missing entry from the posterior predictive distribution:

Round 1:

Student	$Y(1)$	$Y(0)$	τ_i
A	85	drew 72	13
B	78	drew 69	9
C	92	drew 74	18
D	drew 88	70	18
E	drew 82	74	8
F	drew 91	68	23

$\hat{\tau}^{(1)} = (13 + 9 + 18 + 18 + 8 + 23)/6 = 14.83$

Round 2:

$\hat{\tau}^{(2)} = 12.50$ (different random draws)

Round 3:

$\hat{\tau}^{(3)} = 16.17$

Round 4:

$\hat{\tau}^{(4)} = 13.33$

Round 5:

$\hat{\tau}^{(5)} = 15.67$

(In practice, repeat 1000+ times)

Step 4: Summarize posterior distribution

```
ATE Posterior Distribution (illustration):

Frequency
  |
  |
  |
  |
  |
  +------------------→ ATE
  8    12   14   18   22
          ↑
       Mean   14.5
```

- **Posterior mean** 14.5 (point estimate)
- **95% Credible Interval**: Take the 2.5% and 97.5% quantiles, e.g., [7, 22]

Comparison with Neyman's approach

	Neyman	Bayesian
Point estimate	14.33	14.5
Interval	[5.69, 22.97] (confidence interval)	[7, 22] (credible interval)
Interpretation	95% reliability of the method	95% probability that the true ATE lies in this range

The Bayesian credible interval allows us to directly say "there is a 95% probability that the true effect is in this range" – an interpretation that the frequentist confidence interval **cannot** support.

Example 5: Bayesian Bootstrap

Goal

Estimate the distribution of the ATE using a non-parametric method, without assuming any distributional form.

Step 1: Determine the sampling pools

- Missing $Y_i(0)$ (treated units A, B, C) → draw from observed control values **{70, 74, 68}**
- Missing $Y_i(1)$ (control units D, E, F) → draw from observed treatment values **{85, 78, 92}**

Step 2: Full enumeration

6 missing values, each with 3 possibilities → $3^6 = 729$ combinations. Too many, so we illustrate a few rounds:

Round 1:

```
A's Y(0): draw from {70,74,68} → 74
B's Y(0): draw from {70,74,68} → 68
C's Y(0): draw from {70,74,68} → 70
D's Y(1): draw from {85,78,92} → 85
E's Y(1): draw from {85,78,92} → 92
F's Y(1): draw from {85,78,92} → 78
```

Student	$Y(1)$	$Y(0)$	τ_i
A	85	74	11
B	78	68	10
C	92	70	22
D	85	70	15
E	92	74	18
F	78	68	10

$\hat{\tau}^{(1)} = (11 + 10 + 22 + 15 + 18 + 10)/6 = 14.33$

Round 2:

```
A → 70, B → 70, C → 74, D → 92, E → 78, F → 85
```

Student	$Y(1)$	$Y(0)$	τ_i
A	85	70	15
B	78	70	8
C	92	74	18
D	92	70	22
E	78	74	4
F	85	68	17

$\hat{\tau}^{(2)} = (15 + 8 + 18 + 22 + 4 + 17)/6 = 14.00$

Round 3:

```
A → 68, B → 74, C → 68, D → 78, E → 85, F → 92
```

$\hat{\tau}^{(3)} = (17 + 4 + 24 + 8 + 11 + 24)/6 = 14.67$

Step 3: Repeat 1000 times, summarize

```
ATE Distribution (illustration):

Frequency
  |
  |
  |
  |
  |
  +---------------------→ ATE
  6    10   14   18   22
            ↑
       Mean   14.3
```

- **Point estimate** (mean) 14.3 (very close to $T^{\text{obs}} = 14.33$)
- **95% interval**: approximately [8, 21]

Comparison with Parametric Bayesian Method

	Parametric Multiple Imputation	Bayesian Bootstrap
Assumptions	Required (e.g., normal distribution)	Not required
Sampling source	Assumed distribution $N(\hat{\mu}, \hat{\sigma}^2)$	Actual observed values
Risk	Wrong distributional assumption → bias	Sampling pool too small → coarse distribution
Best for	When a specific distribution is justified	When unsure about distributional shape

In this example, the control group has only 3 values {70, 74, 68}, so the Bootstrap sampling pool is very small and the ATE distribution will be "grainy." But with 50+ units per group, the Bootstrap performs very well.

Appendix: Five Methods Quick Reference

```
                    Experimental Data (Observations + Assignments)
```

```
                 ↓               ↓               ↓

            Fisher         Neyman         Bayesian



      ↓                 ↓     ↓          ↓             ↓
  Test            Conf Int  Est + CI  Parametric    Bootstrap
  sharp null     (inversion) (formula) (mult imp)   (non-param)

  p-value      Fisher CI  Neyman CI  Posterior    Posterior

       → HL Est ←
      (pairwise median)

                       ˆ ± 1.96×SE
```

Method	Input	Output	Sample Size Req.	Assumptions
Fisher Test	Data + sharp null	p-value	Any (but too small = low p-value “resolution”)	Sharp null
HL Estimator	Data	Point estimate $\hat{\tau}_{HL}$	Any	Constant treatment effect
Fisher CI	Data + multiple τ_0	Exact CI	Any	Constant treatment effect
Neyman	Data	$\hat{\tau}$, SE, CI	Large (CLT)	None (allows heterogeneous effects)
Bayesian (Parametric)	Data + model + prior	Posterior distribution	Any	Model correctness
Bayesian Bootstrap	Data	Posterior distribution	Medium (sampling pool must be large enough)	CRD + no parametric assumptions

---

Good luck on the exam!