Faculty Development Programme
ON
Artificial Intelligence and Data Science:
Foundations, Pedagogy, Tools and Emerging Research Trends

RESEARCH METHODOLOGIES IN DATA SCIENCE: HYPOTHESIS TESTING, EXPERIMENT DESIGN, AND PUBLICATION STRATEGIES

Understanding Research Methodology through the Art of Monastic Debate

Dr. Sudhakar Mishra
Asst. Professor
Department of Artificial Intelligence
SVNIT, Surat

The Monastic Debate

The Practice: A dynamic dialogue rooted in ancient traditions (Nalanda/Tibetan).

The Goal: Not to "win" in the Western sense, but to jointly uncover inconsistencies in a philosophical position.

The Method: Active reasoning, rigorous logic, and the exposing of contradictions to reach a deeper truth.

Mapping the Metaphor

The Defender (H_0)

Maintains a consistent philosophical position.

Represents the "Status Quo" or the default assumption.

"All phenomena are permanent."

The Challenger (H_1)

Attempts to find a flaw or contradiction.

Represents the "New Discovery" or the effect we want to prove.

"But a seed changes into a sprout?"

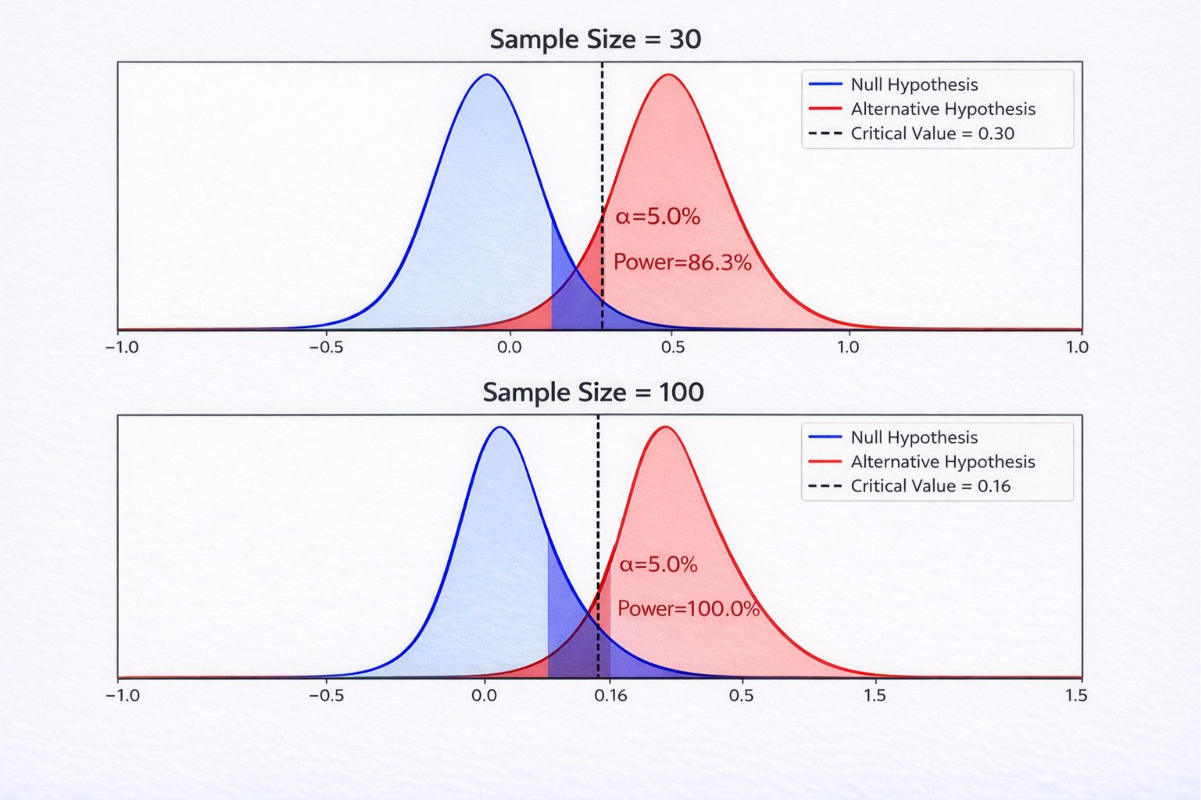
The Null Hypothesis (H_0)

The **Null Hypothesis** is the assumption of "No Difference" or "Consistency".

In Data Science: "The new algorithm performs the same as the old one."

In Debate: "The Defender's logic is sound and contains no contradictions."

We assume H_0 is true until proven otherwise.



The Alternative Hypothesis (H_1)

The **Alternative Hypothesis** is what we are trying to demonstrate.

In Data Science: "The new algorithm has higher accuracy than the baseline."

In Debate: "The Defender's position leads to a logical contradiction."

Burden of Proof

The burden lies entirely on the Challenger (H_1).

Data as Evidence

In Debate

The "Data" consists of the sequence of questions and answers. The Challenger extracts admissions from the Defender.

"You agreed X, but X implies Y, and Y contradicts Z!"

In Data Science

The "Data" consists of our sample observations.

"We observed a 5% increase in conversion rate over 10,000 users."

Type I Error (α)

The False Accusation

Rejecting the Null Hypothesis when it is actually True.

The probability of committing a Type I error equals the significance level (alpha, α)

Debate Context: The Challenger claims to have found a contradiction, but the Defender was actually consistent (the Challenger misunderstood or twisted words).

Consequence: We accept a false discovery.

The Debate Outcome Matrix

	Reality: Defender is Right (H_0 True)	Reality: Defender is Wrong (H_0 False)
DECISION: Reject H_0	Type I Error ($ \alpha$) False Accusation "Seeing a flaw that isn't there"	Correct Decision Valid Refutation (Power)
DECISION: Fail to Reject H_0	Correct Decision Valid Consistency	Type II Error ($ \beta$) Missed Flaw "Failing to see the error"

Type II Error (β)

The Missed Opportunity

Failing to Reject the Null Hypothesis when it is actually False.

Debate Context: The Defender holds a flawed view, but the Challenger is not skilled enough to expose it. The flaw remains hidden.

Consequence: We fail to discover a real effect.

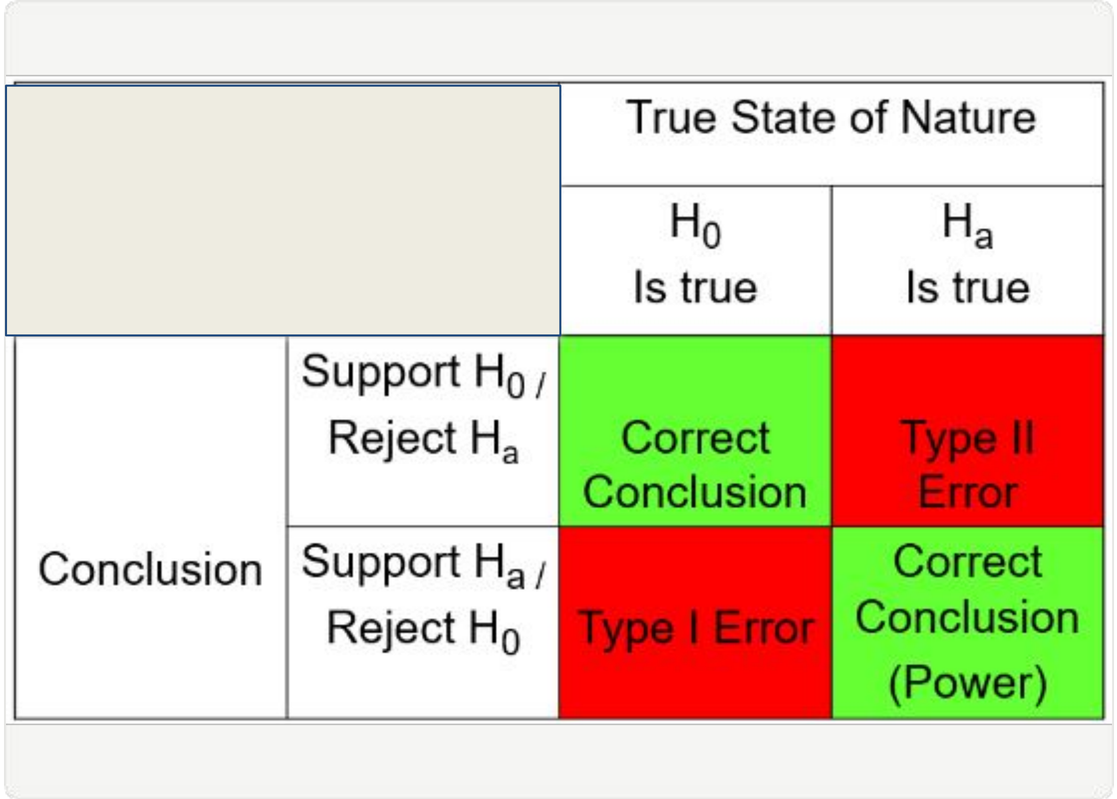
The Debate Outcome Matrix

	Reality: Defender is Right (H_0 True)	Reality: Defender is Wrong (H_0 False)
DECISION: Reject H_0	Type I Error ($ \alpha$) **False Accusation** "Seeing a flaw that isn't there"	Correct Decision **Valid Refutation** (Power)
DECISION: Fail to Reject H_0	Correct Decision **Valid Consistency**	Type II Error ($ \beta$) **Missed Flaw** "Failing to see the error"

The Error Matrix

This table summarizes the four possible outcomes of any hypothesis test or debate conclusion.

- **True Positive:** Correctly identifying a flaw.
- **True Negative:** Correctly agreeing the logic is sound.



		True State of Nature	
		H_0 Is true	H_a Is true
Conclusion	Support H_0 / Reject H_a	Correct Conclusion	Type II Error
	Support H_a / Reject H_0	Type I Error	Correct Conclusion (Power)

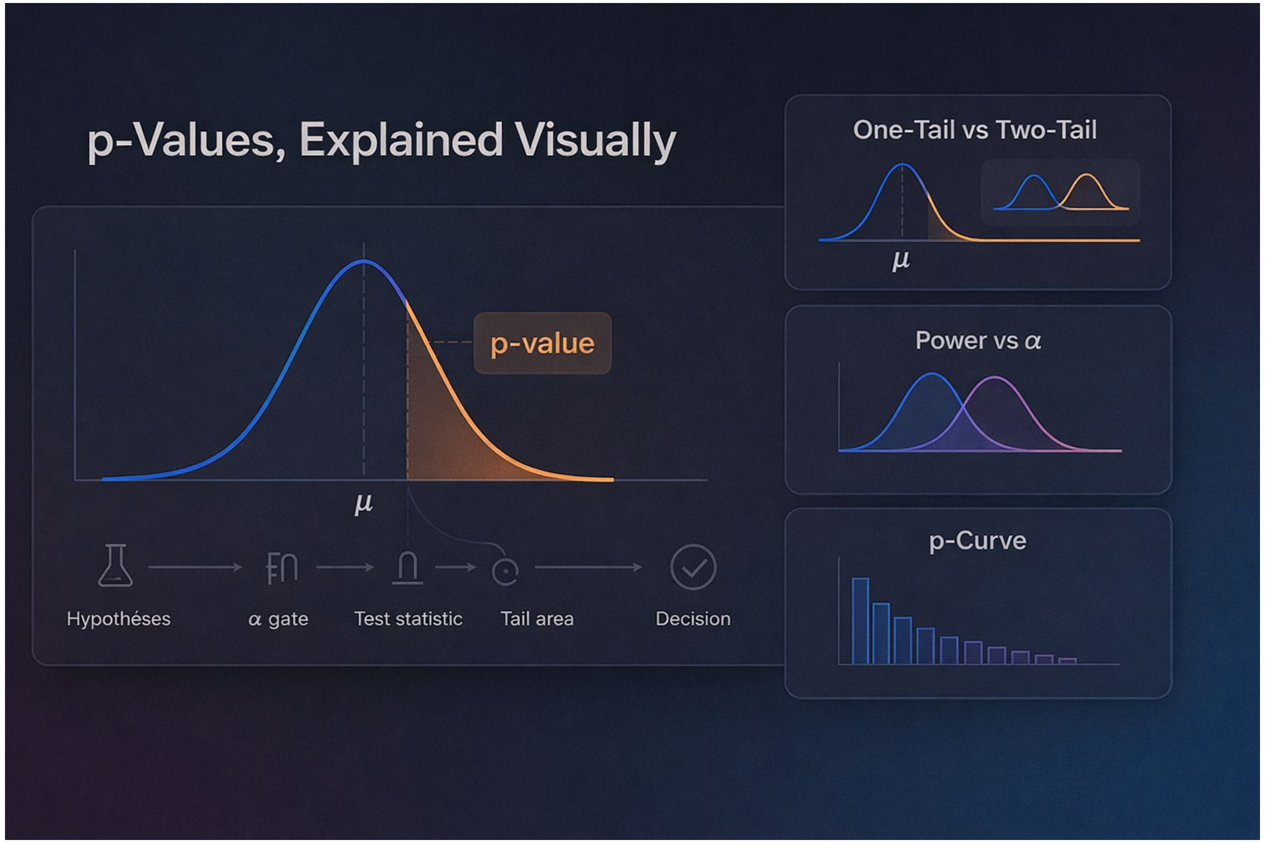
The P-Value

Understanding Probability

The probability of observing the data (evidence) *assuming the Null Hypothesis is true.*

Debate Metaphor: "If the Defender is truly logical (H_0), what are the odds they would accidentally say something this contradictory?"

Low P-Value: "It is highly unlikely a logical person would say this. They must be wrong." ($Reject H_0$)



Significance Level (α)

The "Rules of Debate"

How strictly do we judge the Defender? Usually set at 0.05 (5%).

We accept a 5% risk of making a Type I Error (False Accusation).

Setting the Bar

If we set the bar too high ($\alpha = 0.0001$), the Challenger will almost never win, even if the Defender is wrong (Low Power).

If we set it too low ($\alpha = 0.20$), we will constantly accuse innocent Defenders of being wrong.

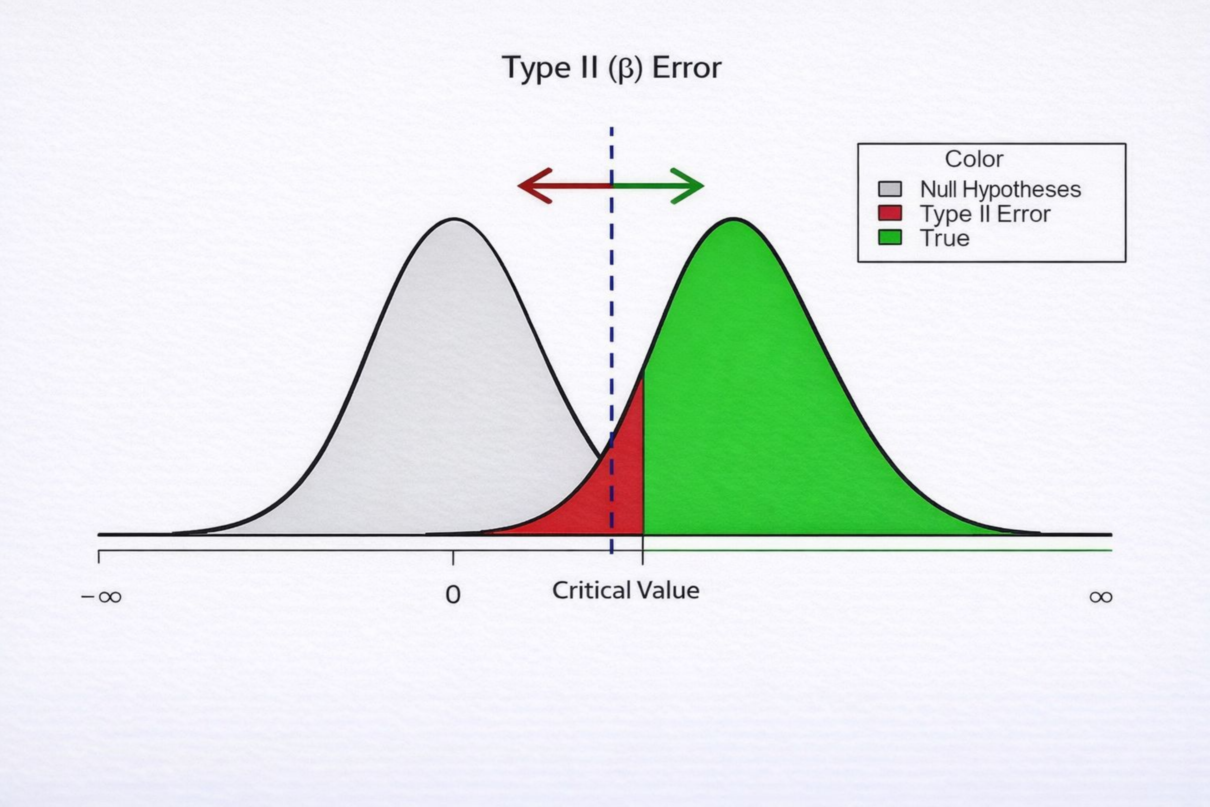
STATISTICAL POWER $(1 - \beta)$

THE SKILL OF THE CHALLENGER

Power is the probability of correctly rejecting a false Null Hypothesis.

In Debate: This corresponds to the Challenger's skill in "Active Reasoning". Can they spot the flaw? Can they formulate the right questions?

Higher sample size (more questions) = Higher Power.



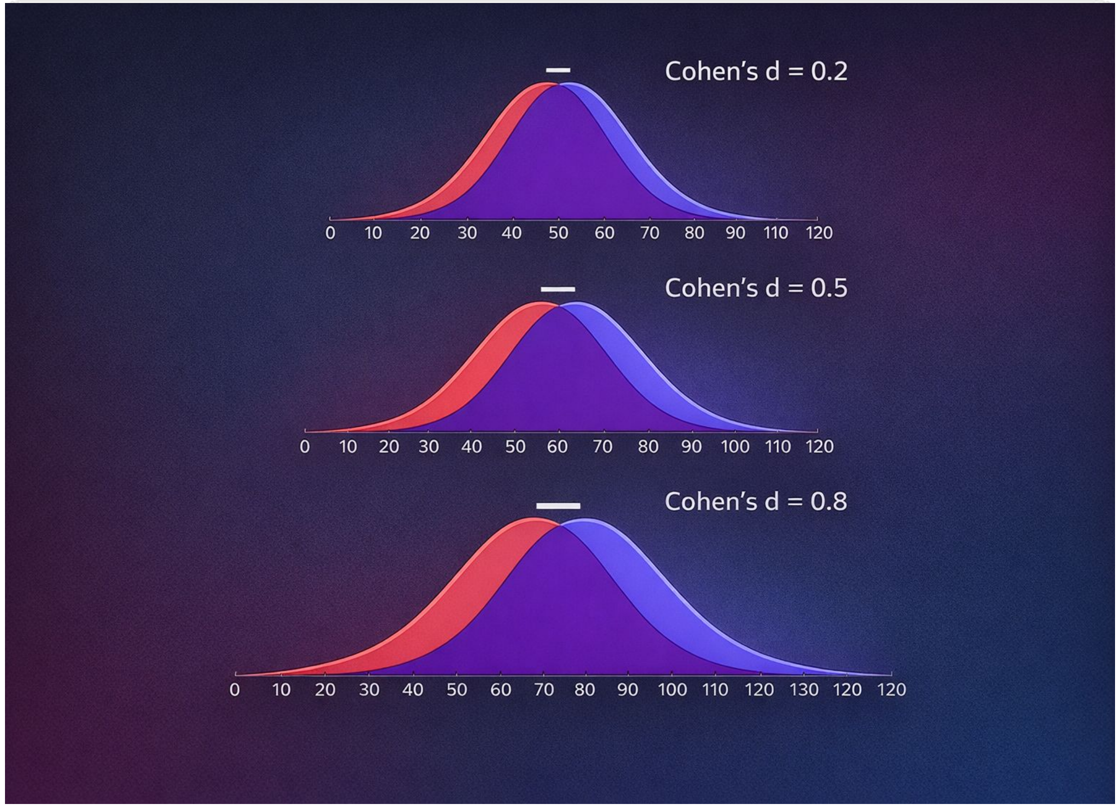
Effect Size

Trivial vs. Fatal Flaws

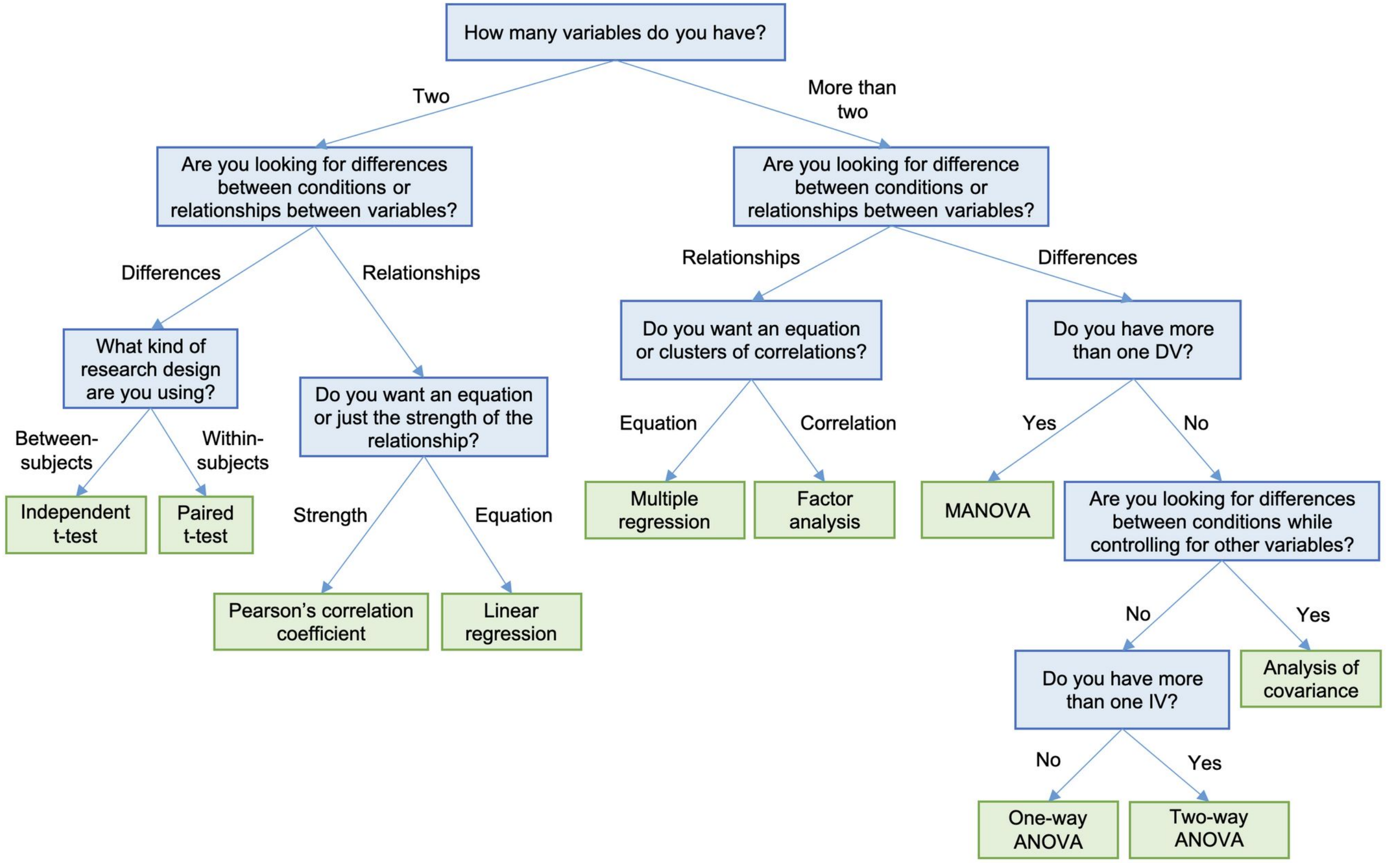
Statistical Significance tells us "Is there a difference?"

Effect Size tells us "How big is the difference?"

In Debate: Did the Defender make a tiny grammatical slip (Low Effect Size) or did they fundamentally contradict their core philosophy (High Effect Size)?



Taxonomy of Hypothesis Testing



How many variables do you have?

Two

More than two

Are you looking for differences between conditions or relationships between variables?

Differences

Relationships

What kind of research design are you using?

Between-subjects

Within-subjects

Independent t-test

Paired t-test

Do you want an equation or just the strength of the relationship?

Strength

Equation

Pearson's correlation coefficient

Linear regression

Are you looking for difference between conditions or relationships between variables?

Relationships

Differences

Do you want an equation or clusters of correlations?

Equation

Correlation

Multiple regression

Factor analysis

Do you have more than one DV?

Yes

No

MANOVA

Are you looking for differences between conditions while controlling for other variables?

No

Yes

Analysis of covariance

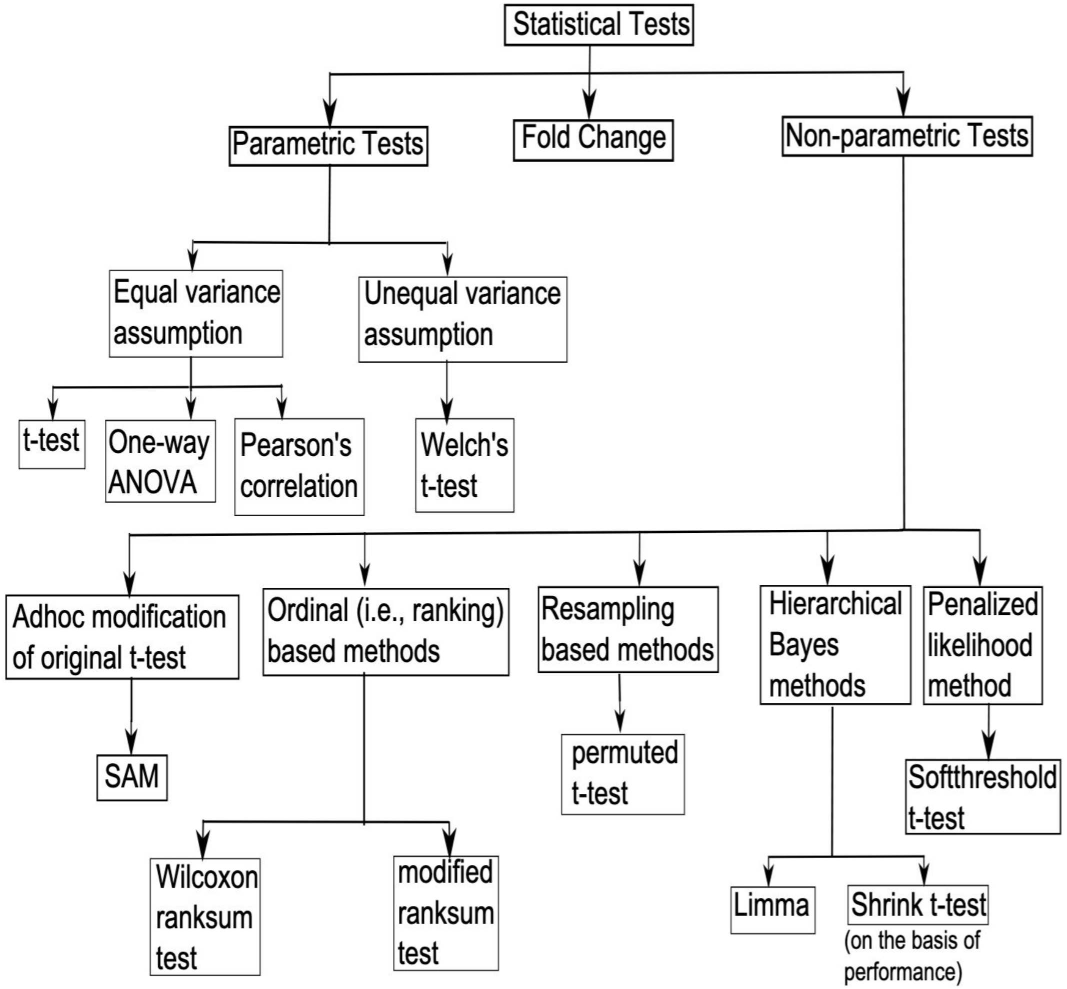
Do you have more than one IV?

No

Yes

One-way ANOVA

Two-way ANOVA



Statistical Tests

Parametric Tests

Fold Change

Non-parametric Tests

Equal variance assumption

Unequal variance assumption

t-test

One-way ANOVA

Pearson's correlation

Welch's t-test

Adhoc modification of original t-test

Ordinal (i.e., ranking) based methods

Resampling based methods

Hierarchical Bayes methods

Penalized likelihood method

SAM

Wilcoxon ranksum test

modified ranksum test

permuted t-test

Limma

Shrink t-test (on the basis of performance)

Softthreshold t-test

Issues with Traditional p-Value Thresholds

- Arbitrary significance level (0.05) has historical rather than mathematical justification
- Threshold creates a "cliff effect" where $p = 0.051$ and $p = 0.049$ are treated dramatically differently
- Over-emphasis on statistical significance rather than practical significance
- Incentivizes p-hacking and questionable research practices

Can we do anything about it?

Issues with Traditional p-Value Thresholds

- Arbitrary significance level (0.05) has historical rather than mathematical justification
- Threshold creates a "cliff effect" where p = 0.051 and p = 0.049 are treated dramatically differently
- Over-emphasis on statistical significance rather than practical significance
- Incentivizes p-hacking and questionable research practices

Recent Recommendations:

- Leading statisticians have proposed more stringent thresholds (p < 0.005) for novel claims to reduce false discovery rates.
- However, p-values should always be reported alongside effect sizes and confidence intervals.
- Bayesian Statistics to get the evidence for alternative hypotheses.

A Paradigm Shift

From "Refuting" to "Updating Beliefs"

Entering the Bayesian Perspective

The Bayesian Prior

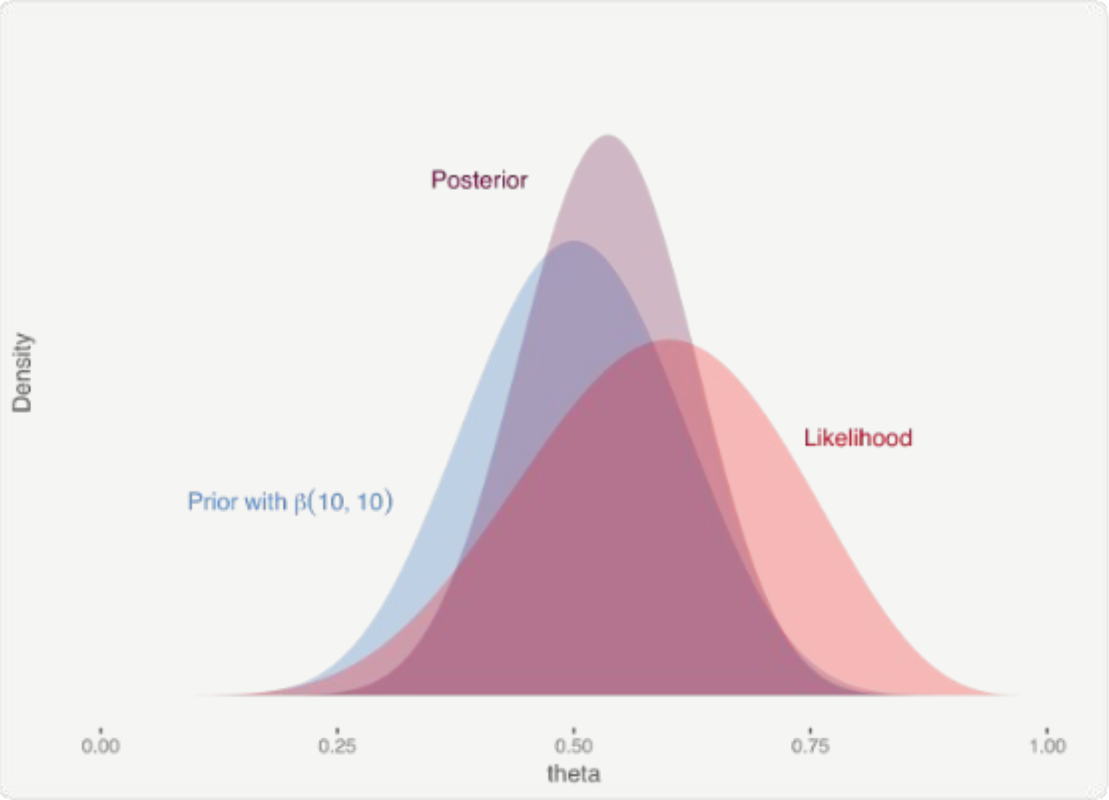
Bayesian

Pre-existing Beliefs

In Frequentist testing, we start blank. In Bayesian, we start with a **Prior**.

Debate Analogy: Before the debate starts, how much do we trust the Defender's wisdom? Is this a novice monk (Weak Prior) or the Dalai Lama (Strong Prior)?

"Extraordinary claims require extraordinary evidence."



The Likelihood

The Debate Itself

This represents the new evidence gathered during the debate.

How likely is this specific exchange of arguments given the Defender is right? vs. given they are wrong?

Data Weight

A long, rigorous debate (lots of data) has a sharper likelihood function. It provides strong evidence that can overwhelm the Prior.

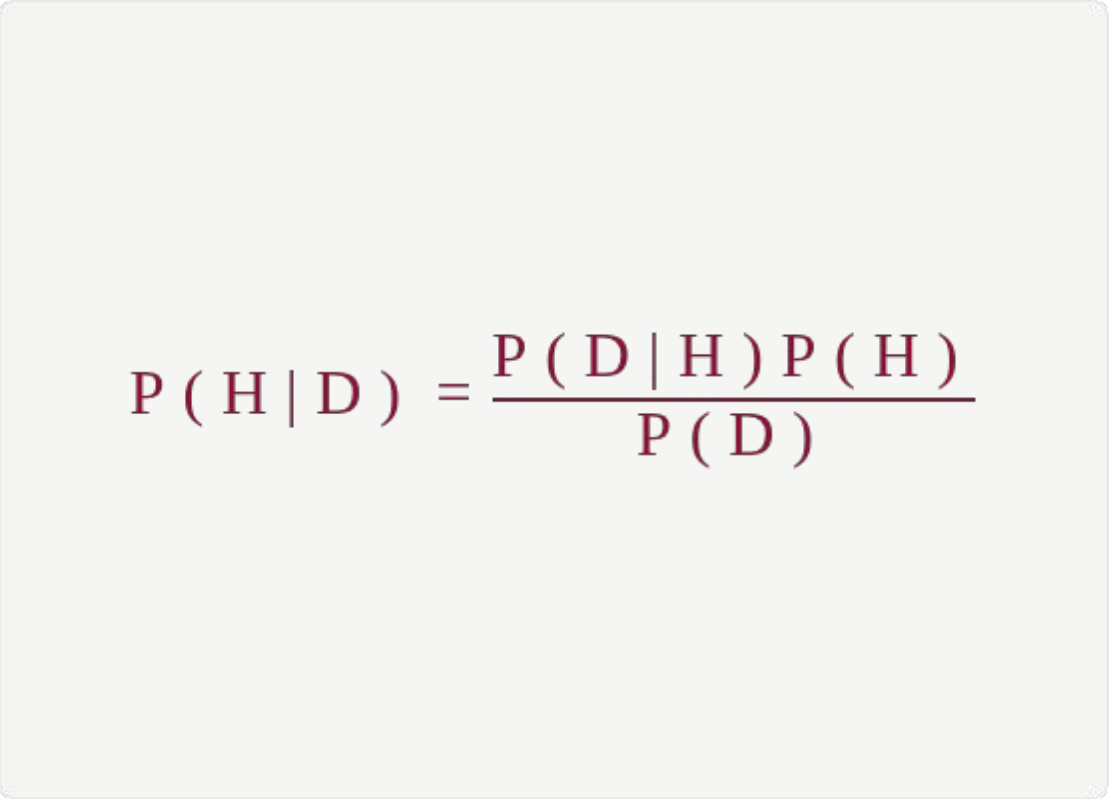
The Posterior

The Updated Belief

Prior × Likelihood ∝ Posterior

After hearing the debate, what do we believe now?

If we had a Strong Prior (Dalai Lama) and weak evidence, our belief barely changes. If the evidence is overwhelming, even a Strong Prior shifts.



$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Bayes Factor

Quantifying the Winner

A ratio comparing the predictive power of two competing hypotheses.

$$BF_{01} = \frac{data/H_0}{data/H_1}$$

p-Value

H_1 : Evidence favors the Challenger ($BF > 1$).

H_0 : Evidence favors the Defender ($BF < 1$).

Unlike P-values, this allows us to gather evidence **in favor** of the Null or alternative..

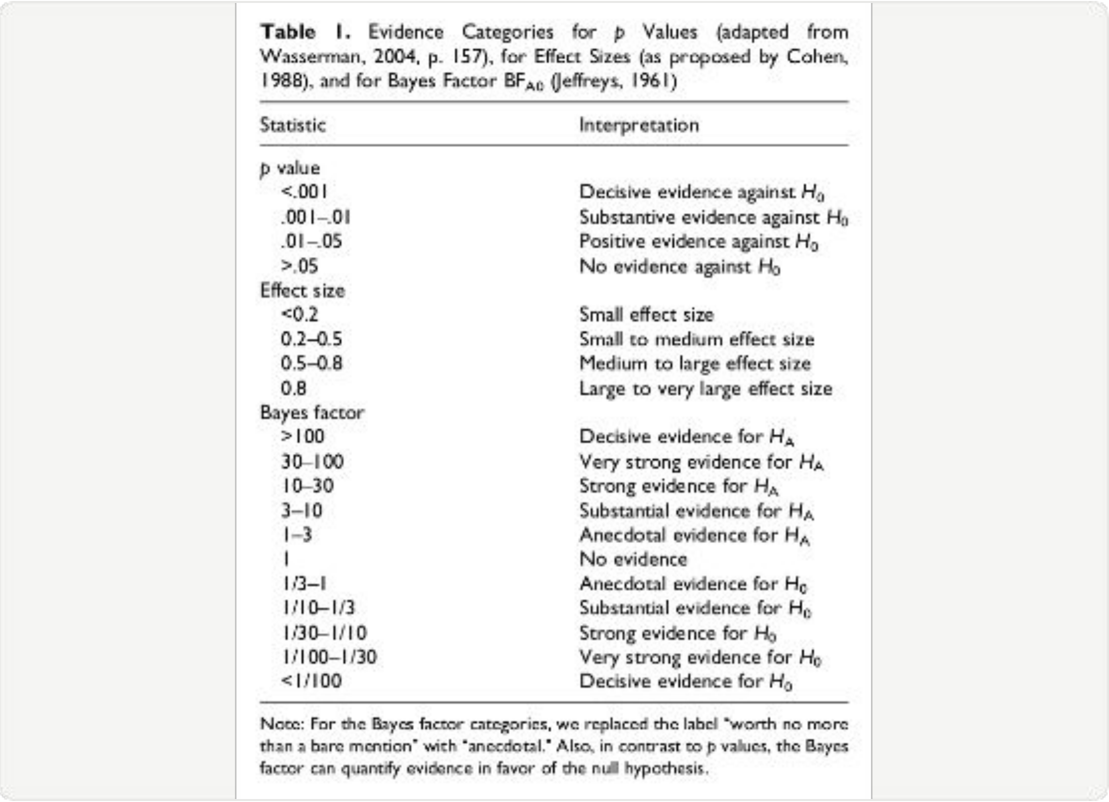


Table 1. Evidence Categories for *p* Values (adapted from Wasserman, 2004, p. 157), for Effect Sizes (as proposed by Cohen, 1988), and for Bayes Factor BF_{A0} (Jeffreys, 1961)

Statistic	Interpretation
p value	
<.001	Decisive evidence against H_0
.001–.01	Substantive evidence against H_0
.01–.05	Positive evidence against H_0
>.05	No evidence against H_0
Effect size	
<0.2	Small effect size
0.2–0.5	Small to medium effect size
0.5–0.8	Medium to large effect size
0.8	Large to very large effect size
Bayes factor	
>100	Decisive evidence for H_A
30–100	Very strong evidence for H_A
10–30	Strong evidence for H_A
3–10	Substantial evidence for H_A
1–3	Anecdotal evidence for H_A
1	No evidence
1/3–1	Anecdotal evidence for H_0
1/10–1/3	Substantial evidence for H_0
1/30–1/10	Strong evidence for H_0
1/100–1/30	Very strong evidence for H_0
<1/100	Decisive evidence for H_0

Note: For the Bayes factor categories, we replaced the label "worth no more than a bare mention" with "anecdotal." Also, in contrast to *p* values, the Bayes factor can quantify evidence in favor of the null hypothesis.

Confidence vs. Credible Intervals

Confidence Interval (Frequentist)

"If we repeated this debate 100 times, 95 of the intervals constructed would contain the true logic."

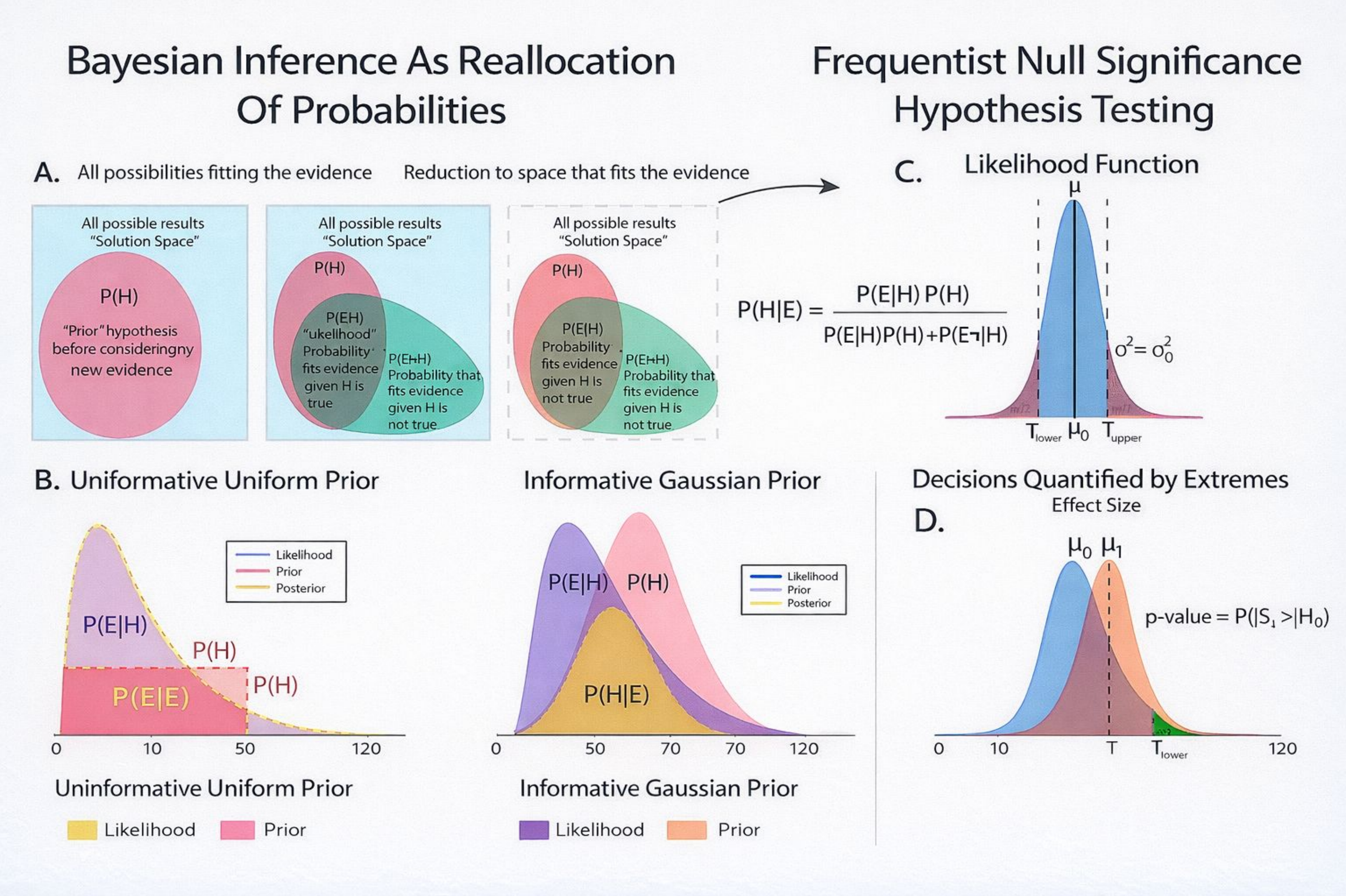
(Counter-intuitive definition).

Credible Interval (Bayesian)

"There is a 95% probability that the Defender's logic falls within this range."

(Intuitive definition).

Bayesian vs Frequentists


Bayesian Inference As Reallocation Of Probabilities
Frequentist Null Significance Hypothesis Testing
A.
All possibilities fitting the evidence
Reduction to space that fits the evidence
All possible results "Solution Space"
P(H)
"Prior" hypothesis before consideringny new evidence
All possible results "Solution Space"
P(H)
P(EH)
"ukellhood"
Probability
fits evidence given H is true
P(E⊢H)
Probability that fits evidence given H is not true
All possible results "Solution Space"
P(H)
P(E|H)
Probability fits evidence given H is not true
P(E⊢H)
Probability that fits evidence given H is not true
$P(H|E) = \frac{P(E|H)\,P(H)}{P(E|H)P(H)+P(E\neg|H)}$
C.
Likelihood Function
μ
$o^2 = o_0^2$
T_{lower}
μ_0
T_{upper}
B. Uniformative Uniform Prior
Likelihood
Prior
Posterior
P(E|H)
P(H)
P(H)
P(E|E)
0
10
50
120
Uninformative Uniform Prior
Likelihood
Prior
Informative Gaussian Prior
P(E|H)
P(H)
Likelihood
Prior
Posterior
P(H|E)
0
50
70
70
120
Informative Gaussian Prior
Likelihood
Prior
Decisions Quantified by Extremes
Effect Size
D.
μ_0
μ_1
p-value = $P(|S_1 >|H_0)$
0
10
T
T_{lower}
120

Conclusion

- **Frequentist (P -value):** Testing the Challenger's ability to refute the Defender. Focus on error rates.
- **Bayesian:** Updating our trust in the Defender based on new evidence. Focus on probability of truth.
- **Monastic Debate:** Both are forms of "Active Reasoning" designed to peel away layers of confusion and arrive at the truth.

"Insight comes from the clash of differing views."

Experimental Design in Data Science

Illustrated through the lens of Monastic Debate

Concepts: RCT, Factorial, Quasi-Experiments, and Bias

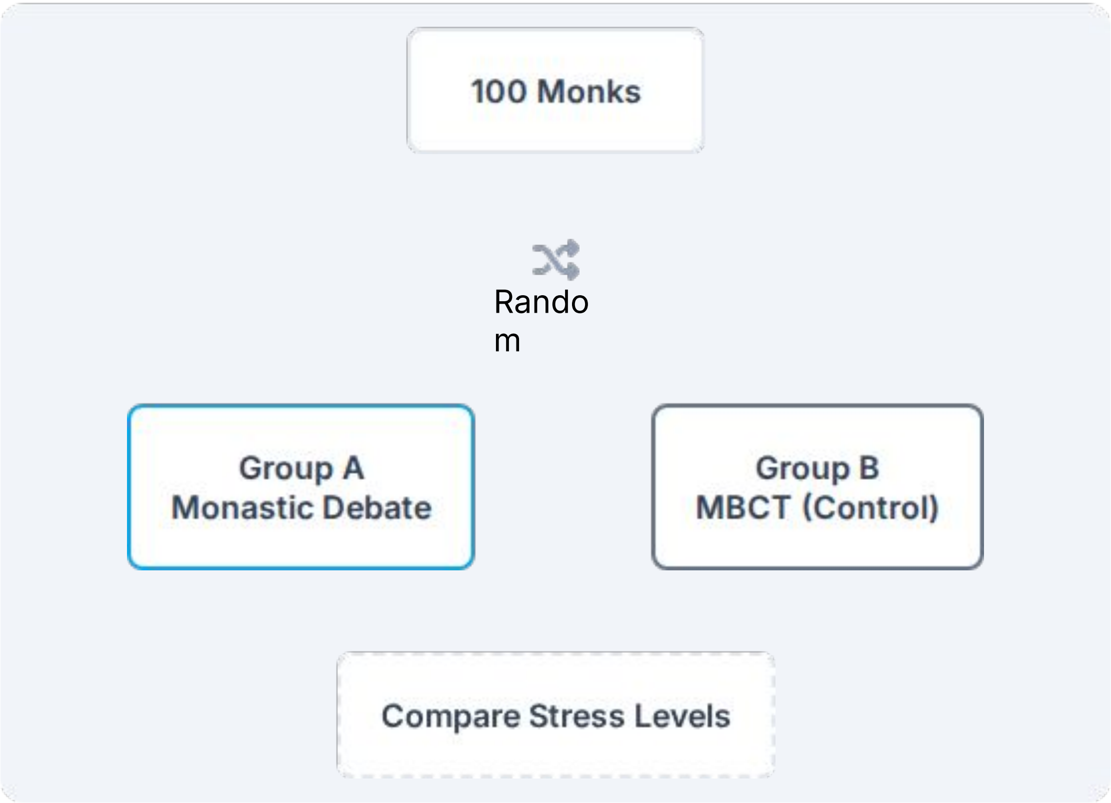
1. Randomized Controlled Trial (RCT)

The Gold Standard

Concept: Randomly assigning subjects to "Treatment" and "Control" groups to eliminate selection bias.

Debate Example: To test if *Monastic Debate* improves **Emotion Regulation** better than *MBCT*:

- **Population:** 100 Novice Monks.
- **Randomization:** Coin flip assigns 50 to Debate, 50 to MBCT.
- **Measurement:** Stress response after 6 months.



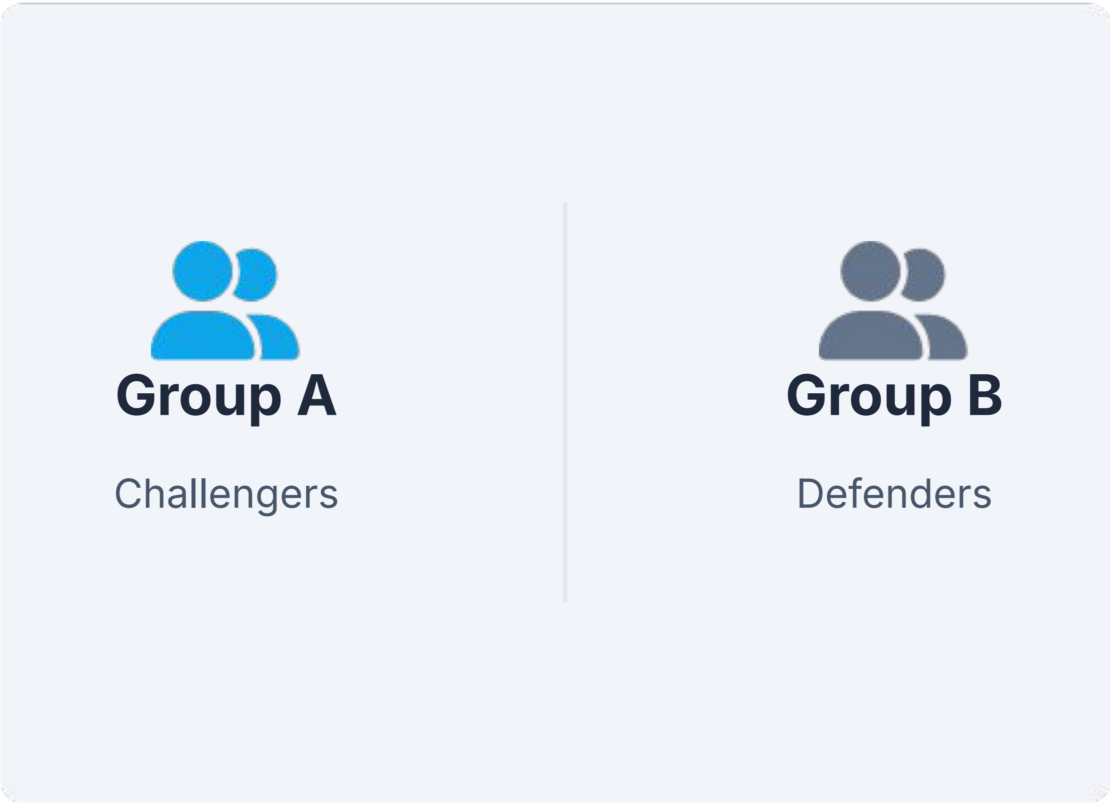
2. Between-Subjects Design

Distinct Groups, Distinct Treatments

Concept: Each participant experiences only *one* condition. Used when one condition influences the other (carryover effects).

Debate Example: Investigating the specific benefits of being a **Defender** vs. a **Challenger**.

- **Group A:** Only acts as Challengers (Active questioning).
- **Group B:** Only acts as Defenders (maintaining consistency).
- **Outcome:** Measure "Cognitive Flexibility" scores.



Group A

Challengers

Group B

Defenders

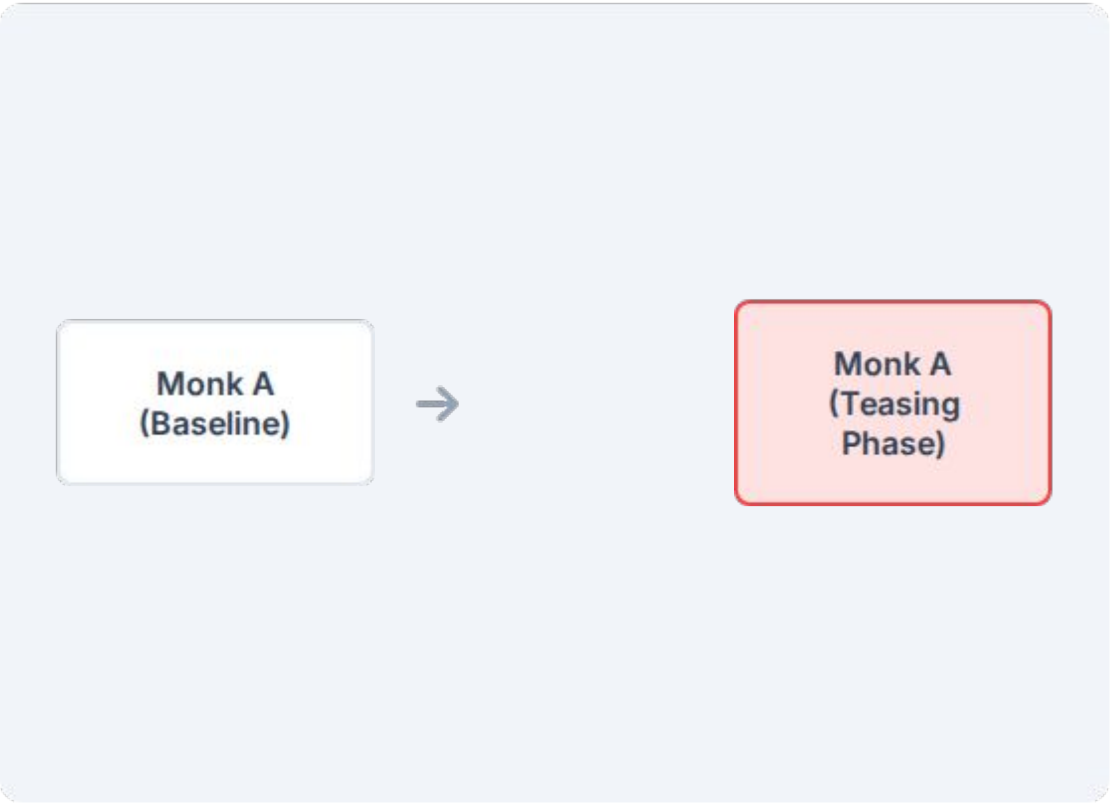
3. Within-Subjects Design

Pre-Post / Repeated Measures

Concept: The same participants experience all conditions. Reduces variance caused by individual differences.

Debate Example: Measuring the *immediate* physiological impact of "Teasing" during debate.

- **Step 1:** Measure Monk A's Heart Rate (HR) during calm logic phase.
- **Step 2:** Measure Monk A's HR during intense "teasing" phase.
- **Comparison:** HR Change within the same monk.



4. Factorial Design (2×2)

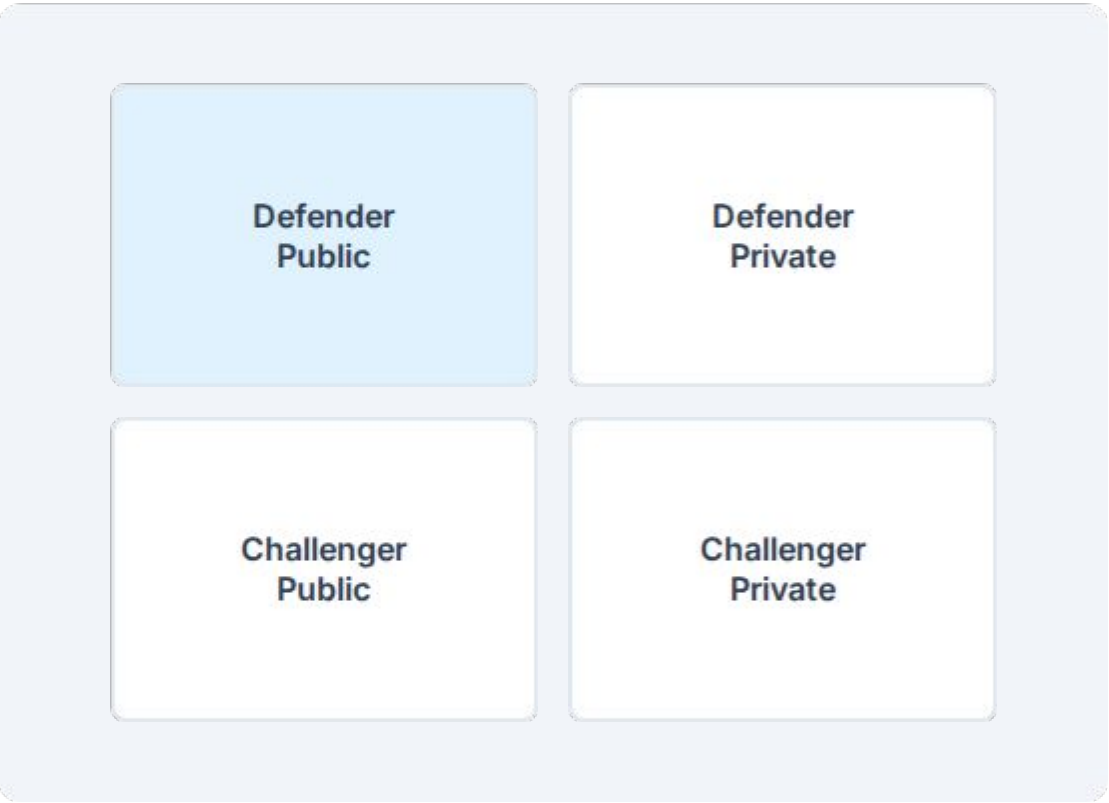
Testing Interactions

Concept: Testing multiple variables (factors) simultaneously to see how they interact.

Debate Example: Factors: **Role** (Challenger/Defender) and **Setting** (Public/Private).

Does the pressure of a *Public* audience affect Defenders more than Challengers?

- Group 1: Defender + Public
- Group 2: Defender + Private
- Group 3: Challenger + Public
- Group 4: Challenger + Private



Defender Public	Defender Private
Challenger Public	Challenger Private

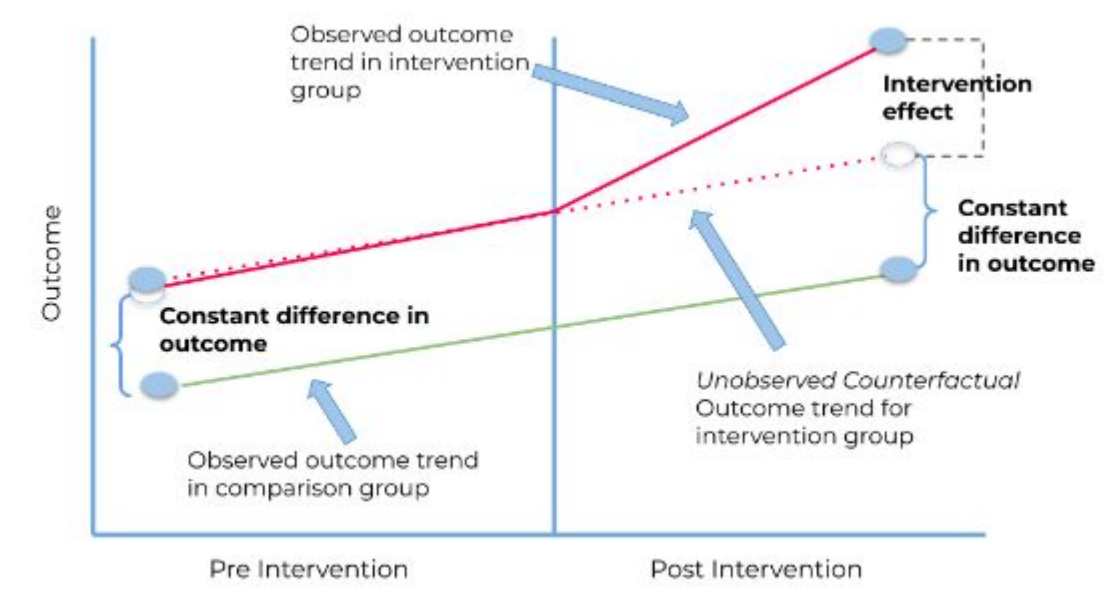
5. Quasi-Experiments (DiD)

No Randomization Possible

Concept: Using existing groups when randomization is unethical or impossible. Often uses "Difference in Differences" (DiD).

Debate Example: We cannot randomly force monks to switch sects.

- **Intervention Group:** Gelug School (Practices Debate).
- **Control Group:** Nyingma School (Practices Meditation only).
- **Method:** Measure the *change* in logic scores over 5 years for both schools and compare the slopes.


Observed outcome trend in intervention group
Intervention effect
Constant difference in outcome
Outcome
Constant difference in outcome
Unobserved Counterfactual Outcome trend for intervention group
Observed outcome trend in comparison group
Pre Intervention
Post Intervention

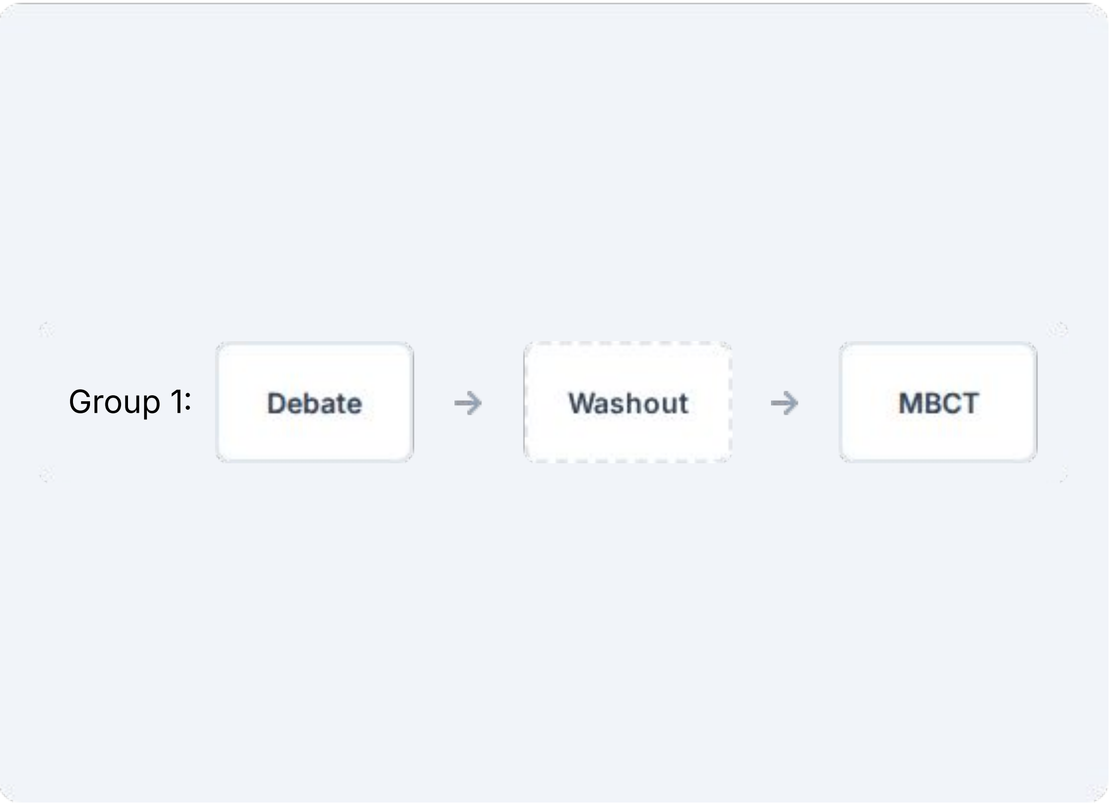
6. Crossover Design

Sequential Treatments

Concept: Participants receive Sequence A then B, or B then A. Requires a "Washout Period".

Debate Example: Does Debate prime the mind for MBCT?

- **Group 1:** 3 Months Debate → Washout → 3 Months MBCT.
- **Group 2:** 3 Months MBCT → Washout → 3 Months Debate.
- **Analysis:** Check if MBCT scores are higher *after* Debate than before.



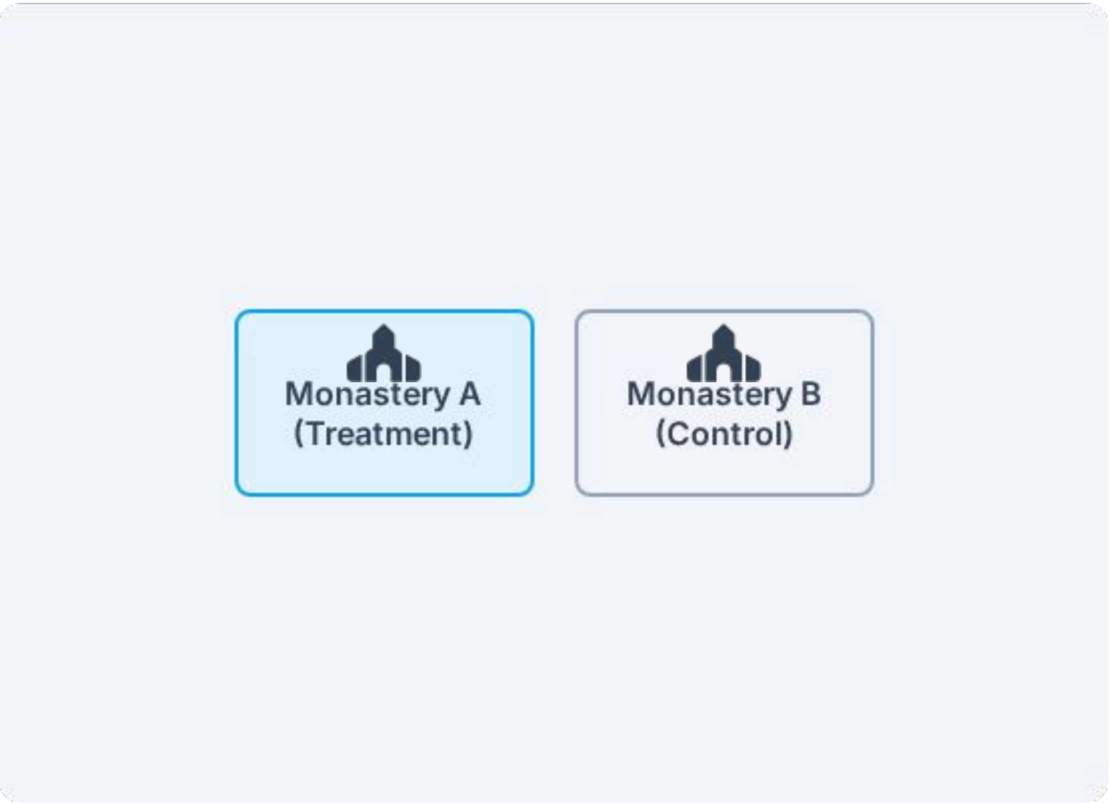
7. Cluster Randomization

Avoiding Contamination

Concept: Randomizing groups (clusters) rather than individuals. Essential when the intervention involves social interaction.

Debate Example: Monastic Debate is social. If we randomize *within* a monastery, "Control" monks will overhear "Treatment" monks debating.

- **Solution:** Randomize *entire monasteries*.
- Monastery A, B, C → Debate Program.
- Monastery D, E, F → Standard Program.



8. Longitudinal (Time Series)

Tracking Changes Over Time

Concept: Repeated observations of the same variables over long periods.

Debate Example: The paper states practitioners "figure out strategies to withstand teasing" over years.

- **Study:** Track a cohort of monks from Novice to Master (10 years).
- **Measure:** Logical consistency and Emotional Reactivity every year.
- **Goal:** Map the learning curve of "Active Reasoning".

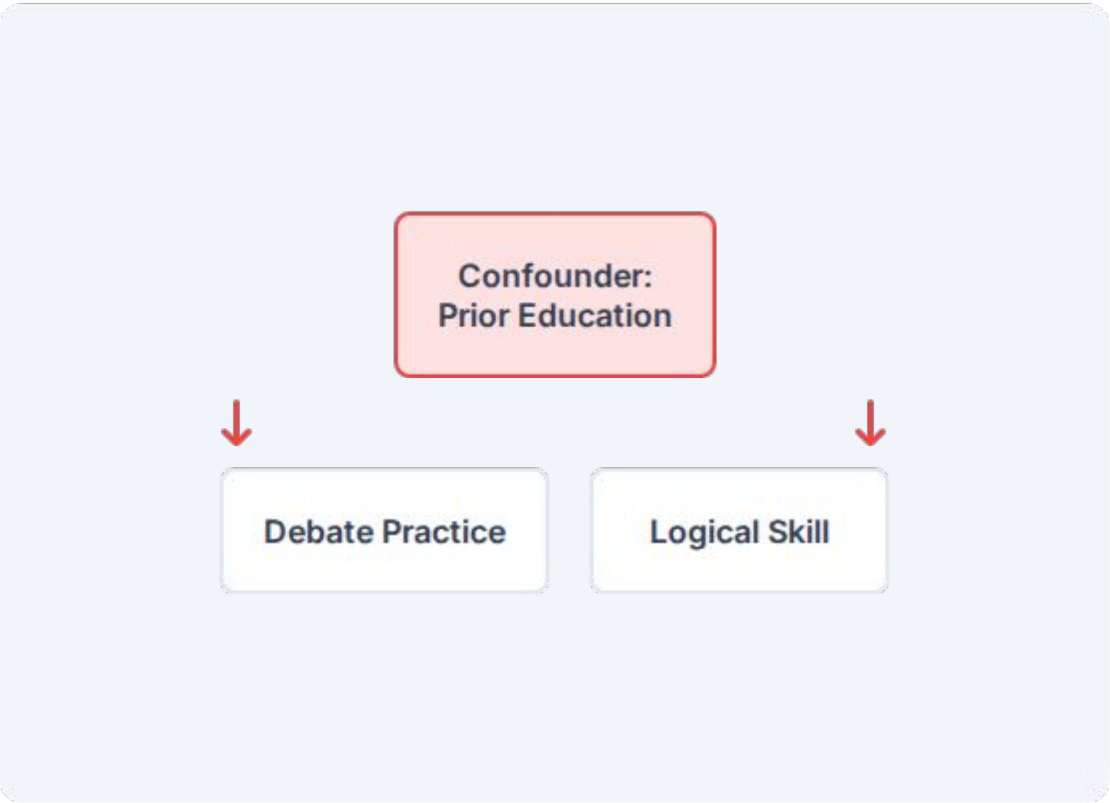
9. Confounding Variables

Threats to Validity

Concept: External factors that correlate with both Independent and Dependent variables, creating false associations.

Debate Example:

- **Observation:** Debating monks have higher IQs.
- **Confounder:** *Selection Bias*. Perhaps smarter novices are encouraged to join the Debate track, while others do chores.
- **Confounder:** *Diet/Lifestyle*. Debate monasteries might have better nutrition.



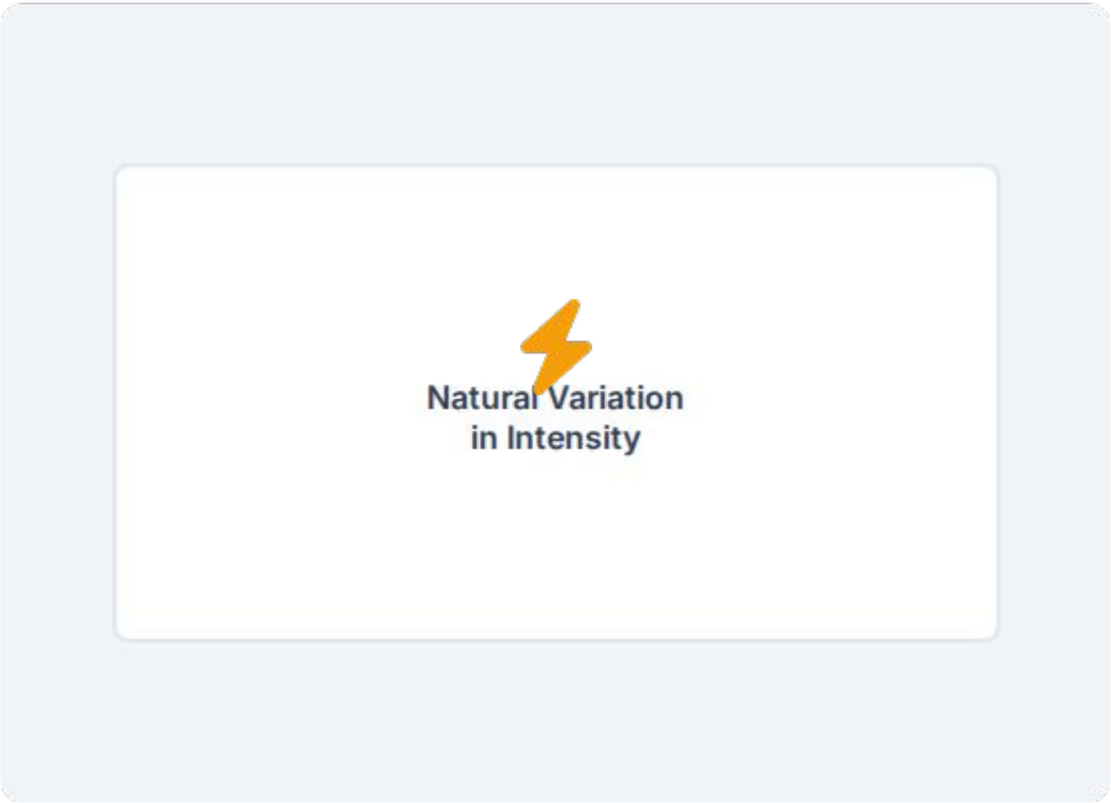
10. Natural Experiments

Exploiting Random Events

Concept: Nature or policy changes create "random" assignments for us.

Debate Example: The "Teasing" (emotional manipulation) varies naturally.

- Some debates naturally become very heated/aggressive due to personality clashes.
- Some debates remain calm.
- **Analysis:** Compare learning outcomes from "High Conflict" vs "Low Conflict" sessions that occurred naturally, controlling for other factors.



Publication Strategies in Data Science

Navigating the Academic Landscape through the Lens of Monastic Debate

1. The Venue: Courtyard vs. Scripture

Conferences (The Courtyard)

Nature: Fast, interactive, public.

Debate Analogy: Like the daily courtyard debates. The goal is rapid exchange of ideas, finding immediate flaws, and real-time interaction (Q&A).

Venues: NeurIPS, ICML, CVPR.

Journals (The Scripture)

Nature: Slow, archival, rigorous.

Debate Analogy: Writing a commentary on the Sutras. It requires deep contemplation, comprehensive references, and perfection of form.

Venues: JMLR, IEEE TPAMI.

2. Targeting the Right Monastery

Select Your Lineage

Not all debates happen in the same school. You must choose where your argument fits.

- 📙 **Gelug School (Logical Rigor):** Equivalent to theoretical venues (COLT: Annual Conference on Learning Theory). Focus on proofs and bounds.
- 📙 **Nyingma School (Practice/Insight):** Equivalent to applied venues (KDD: ACM Transactions on Knowledge Discovery from Data , AAAI: Association for the Advancement of Artificial Intelligence). Focus on utility and real-world application.

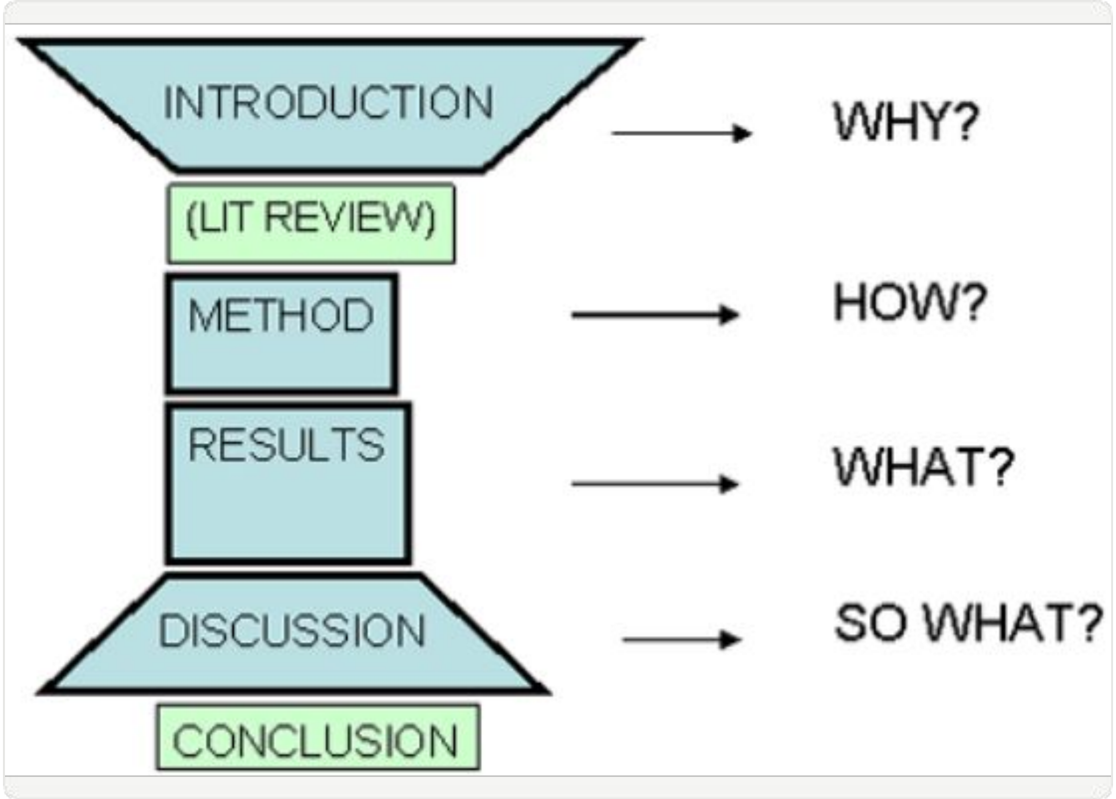
Match the Venue's Scope

3. Structure: The Defender's Stance

In Monastic Debate, the Defender (H_0) must state their position clearly to avoid ambiguity.

IMRaD Structure as Debate Stance:

- **Introduction:** State the Thesis. ("I posit that Transformer X is superior...")
- **Methods:** The Rules of Engagement. How we derived this truth.
- **Results:** The Evidence. The logical consequences of the method.
- **Discussion:** Acknowledging limitations (Self-correction).



4. Peer Review: Enter the Challengers

The Role of Reviewers

Reviewers act as the **Challengers** in a monastic debate.

Their goal is *not* to destroy you, but to find inconsistencies ("logical fallacies") in your work to ensure truth.

"You claimed X implies Y, but in Table 3, Z is observed. This is a contradiction!"

5. The Rebuttal: Active Reasoning

Emotion Regulation

Monastic debate teaches managing emotions under pressure. Do not take reviewer comments personally.

Respond with logic, not defense mechanisms.

Standard Replies

"The reason is not established": The reviewer missed a detail (politely point to line 42).

"I accept": Acknowledge the flaw and fix it. This shows intellectual honesty, a key virtue in debate.

6. Reproducibility: Sharing the Sutras

The Logic Must Hold for All

In debate, a truth must be universal. In Data Science, results must be reproducible.

Artifacts to Publish:

- 📖 **Code:** The "script" of your debate.
- 📖 **Data:** The "evidence" used.
- 📖 **Seeds/Hyperparameters:** The "context" of the argument.

Without code, your paper is just an anecdote.

7. Pre-prints: The Open Courtyard

Rapid Dissemination

Before the formal "examination" (Peer Review), monks often practice in the open courtyard.

ArXiv allows you to stake your claim ("establish priority") and get early feedback from the community.

Risk: No quality filter. You expose your "flawed logic" to the world immediately.

8. Impact: The Lineage

Building on Tradition

In Buddhism, you respect the lineage (previous masters).

In Science, you cite previous work.

High Impact: Your debate clarifies a core confusion, allowing others to build upon it.

Citation Count: A measure of how many other debates rely on your "Defended Position".

9. Ethics: Right Speech

Avoid "P-Hacking"

Torturing data until it confesses is like using rhetorical tricks to win a debate without true insight.

It violates the spirit of the search for truth.

Plagiarism

Reciting another monk's debate as your own.

Always attribute ideas. The goal is collective enlightenment (knowledge), not personal glory.

10. Conclusion

- 📖 **The Paper:** Your Thesis (Defender's Stance).
- 📖 **The Journal:** The Monastery (Venue).
- 📖 **The Reviewers:** The Challengers (Logic Checkers).
- 📖 **The Goal:** Not just to publish, but to contribute a "valid cognition" to the world.

"May this research benefit all sentient beings (and future researchers)."

References

1. Mishra, S., Nusslock, R., Srinivasan, N., & Van Vugt, M. Effect of Contemplative Monastic Debate Practice on Emotion Regulation and Experience.
2. van Vugt, M. K., Soepa, J., Gyaltsen, J., Gyatso, K., Lodroe, T., Aadhentsang, T., ... & Mishra, S. (2023). Using the body to think: an analysis of the cognitive mechanisms underlying Thinking at the Edge and Tibetan monastic debate.
3. van Vugt, M. K., Pollock, J., Johnson, B., Gyatso, K., Norbu, N., Lodroe, T., ... & Fresco, D. M. (2020). Inter-brain synchronization in the practice of Tibetan monastic debate. *Mindfulness*, *11*(5), 1105-1119.
4. Van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., ... & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1.
5. Kirk, R. E. (2009). Experimental design. *Sage handbook of quantitative methods in psychology*, 23-45.
6. Emmert-Streib, F., & Dehmer, M. (2019). Understanding statistical hypothesis testing: The logic of statistical inference. *Machine Learning and Knowledge Extraction*, *1*(3), 945-962.
7. Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, *13*(3), e1002106.