When is statistical evidence strong enough? Using hypothesis tests to value data collection

Aristotle Epanomeritakis Davide Viviano

Department of Economics, Harvard University

When is the evidence enough?

- Growing availability of experiments as inputs for decision-making

  E.g. antipovery programs evaluated by NGOs (GiveWell, 2020), clinical trials FDA approval (Food and Admin, 2019), MVPF for government policies (Hendren and Sprung-Keyser, 2020)

- But when is evidence strong enough to make a decision?

When is the evidence enough?

- Growing availability of experiments as inputs for decision-making

  E.g. antipovery programs evaluated by NGOs (GiveWell, 2020), clinical trials FDA approval (Food and Admin, 2019), MVPF for government policies (Hendren and Sprung-Keyser, 2020)

- But when is evidence strong enough to make a decision?
  - Standard hypothesis testing: implement if the estimated benefit is positive and statistically significant otherwise keep the status quo.
  - Its motivation unclear/debated (Manski, 2021a; Tetenov, 2016)
  - Difficulty comes from two point decisions. But failure to reject can mean "we do not know yet" (Fisher, 1935). So why use size $\alpha = 5\%$?

When is the evidence enough?

- Growing availability of experiments as inputs for decision-making

  E.g. antipovery programs evaluated by NGOs (GiveWell, 2020), clinical trials FDA approval (Food and Admin, 2019), MVPF for government policies (Hendren and Sprung-Keyser, 2020)

- But when is evidence strong enough to make a decision?
  - Standard hypothesis testing: implement if the estimated benefit is positive and statistically significant otherwise keep the status quo.
  - Its motivation unclear/debated (Manski, 2021a; Tetenov, 2016)
  - Difficulty comes from two point decisions. But failure to reject can mean "we do not know yet" (Fisher, 1935). So why use size $\alpha = 5\%$?

- **Goal:** can we quantify how much additional evidence is worth and when instead we should make immediate policy decision?

Summary of the paper

- Three point decisions: implement a policy, reject or "I do not know"
- "I do not know" corresponds to collect more evidence tomorrow at given cost (and with given precision) and then make a decision
- Decision that minimizes worst-case welfare loss (regret) is a t-test
- Threshold depends on cost (and precision) of future experiments

- Invert the threshold to introduce the Abstention value
- A-value denotes indifference cost for saying I do not know
- Provides welfare guarantees for allocating scarse resources
- Welfare (monetary) based notion of strength of economic evidence

Related literature

- Rationale for hypothesis testing and treatment choice (Abadie, 2020; Andrews and Chen, 2025; Bates et al., 2026; Benjamin et al., 2018; Imbens, 2021; Manski, 2021a,b; Tetenov, 2012, 2016; Viviano et al., 2026)
  - Statistical testing difficult to rationalize with two point decisions
  - Here three point decisions allowing for I do not know
  - Critical value endogenous as a function of cost of data collection

Related literature

- Rationale for hypothesis testing and treatment choice (Abadie, 2020; Andrews and Chen, 2025; Bates et al., 2026; Benjamin et al., 2018; Imbens, 2021; Manski, 2021a,b; Tetenov, 2012, 2016; Viviano et al., 2026)
  - Statistical testing difficult to rationalize with two point decisions
  - Here three point decisions allowing for I do not know
  - Critical value endogenous as a function of cost of data collection

- Sequential testing and rejection option (Adusumilli, 2026; Arrow et al., 1949; Breza et al., 2025; Chernoff, 1959; Chow, 1970; Horowitz and Lee, 2025; Manski and Tetenov, 2016; Maurice, 1957; Wald, 1947; Wald and Wolfowitz, 1950)
  - Classical literature gives foundation for the problem
  - Here introduce A-value by establishing monotonicity of threshold
  - Analyze the A-value to allocate resources over multiple experiments

Related literature

- Rationale for hypothesis testing and treatment choice (Abadie, 2020; Andrews and Chen, 2025; Bates et al., 2026; Benjamin et al., 2018; Imbens, 2021; Manski, 2021a,b; Tetenov, 2012, 2016; Viviano et al., 2026)
  - Statistical testing difficult to rationalize with two point decisions
  - Here three point decisions allowing for I do not know
  - Critical value endogenous as a function of cost of data collection

- Sequential testing and rejection option (Adusumilli, 2026; Arrow et al., 1949; Breza et al., 2025; Chernoff, 1959; Chow, 1970; Horowitz and Lee, 2025; Manski and Tetenov, 2016; Maurice, 1957; Wald, 1947; Wald and Wolfowitz, 1950)
  - Classical literature gives foundation for the problem
  - Here introduce A-value by establishing monotonicity of threshold
  - Analyze the A-value to allocate resources over multiple experiments

- Bayesian decisions (Abadie et al., 2023; Azevedo et al., 2020, 2023; Bather, 1962; Fudenberg et al., 2018; Grundy et al., 1956, 1954; Liang et al., 2022)
  - Here study minimax regret instead of using prior
  - Introduce A-value as "frequentist" analog to Bayesian information

Outline

1. Setup
2. Allocating scarse resources
3. Empirical application
4. Conclusion

Outline

1. Setup
2. Allocating scarse resources
3. Empirical application
4. Conclusion

Basic framework

- Initial estimate of the policy's welfare effect:

$$X_1 \sim \mathcal{N}(\Delta, s_1^2).$$

Δ: welfare gain over the status quo; s: standard error.

- Three possible recommendations:

$$\delta_1(X_1) \in \{ \underbrace{1}_{\text{implement}}, \underbrace{0}_{\text{reject}}, \underbrace{\texttt{NA}}_{\text{collect evidence}} \}.$$

Basic framework

- Initial estimate of the policy's welfare effect:

$$X_1 \sim \mathcal{N}(\Delta, s_1^2).$$

  Δ: welfare gain over the status quo; s: standard error.

- Three possible recommendations:

$$\delta_1(X_1) \in \{ \underbrace{1}_{\text{implement}}, \underbrace{0}_{\text{reject}}, \underbrace{\texttt{NA}}_{\text{collect evidence}} \}.$$

- If we abstain, pay cost c and observe

$$X_2 \sim \mathcal{N}(\Delta, s_2^2), \qquad X_2 \perp X_1.$$

  Then make a final decision $\delta_2(X_1, X_2) \in \{0, 1\}$.

- Denote $w = s_2^2/s_1^2$

Decision problem

Let D be the terminal policy decision (either implement or status quo) and $D^\star = 1\{\Delta \geq 0\}$ the optimal decision. Consider a loss

$$\mathcal{L}(\delta; \Delta, c) = \mathbb{E}_\Delta \left[\underbrace{|\Delta| \, 1\{D \neq D^\star\}}_{\text{wrong policy decision}} + \underbrace{c \, 1\{\delta_1 = \texttt{NA}\}}_{\text{further experimentation}} \right].$$

Decision problem

Let D be the terminal policy decision (either implement or status quo) and $D^\star = 1\{\Delta \geq 0\}$ the optimal decision. Consider a loss

$$\mathcal{L}(\delta; \Delta, c) = \mathbb{E}_\Delta \left[\underbrace{|\Delta| \, 1\{D \neq D^\star\}}_{\text{wrong policy decision}} + \underbrace{c \, 1\{\delta_1 = \texttt{NA}\}}_{\text{further experimentation}} \right].$$

Choose the decision rules to minimize worst-case regret:

$$\min_{\delta_1, \delta_2} \sup_{\Delta} \mathcal{L}(\delta; \Delta, c).$$

Decision problem

Let D be the terminal policy decision (either implement or status quo) and $D^\star = 1\{\Delta \geq 0\}$ the optimal decision. Consider a loss

$$\mathcal{L}(\delta; \Delta, c) = \mathbb{E}_\Delta \Big[\underbrace{|\Delta|\, 1\{D \neq D^\star\}}_{\text{wrong policy decision}} + \underbrace{c\, 1\{\delta_1 = \texttt{NA}\}}_{\text{further experimentation}} \Big].$$

Choose the decision rules to minimize worst-case regret:

$$\min_{\delta_1, \delta_2} \sup_{\Delta} \mathcal{L}(\delta; \Delta, c).$$

	Without waiting		After waiting	
True welfare effect	Implement	Reject	Implement	Reject
Beneficial policy: $\Delta > 0$	0	Δ	c	$c + \Delta$
Harmful policy: $\Delta < 0$	$-\Delta$	0	$c - \Delta$	c

(Extension later: weigh minimax solution against a prior over Δ)

Optimal decision: t-test structure

The following is the (almost sure) unique minimax optimal decision

$$\delta_1^\star(X_1) = \begin{cases} 1 & X_1/s \geq \tau^\star(c/s; w), \\ 0 & X_1/s \leq -\tau^\star(c/s; w), \\ \texttt{NA} & |X_1|/s < \tau^\star(c/s; w) \end{cases} \qquad \delta_2^\star(X_1, X_2) = 1\{wX_1 + X_2 \geq 0\}$$

Optimal decision: t-test structure

The following is the (almost sure) unique minimax optimal decision

$$\delta_1^\star(X_1) = \begin{cases} 1 & X_1/s \geq \tau^\star(c/s; w), \\ 0 & X_1/s \leq -\tau^\star(c/s; w), \\ \texttt{NA} & |X_1|/s < \tau^\star(c/s; w) \end{cases} \qquad \delta_2^\star(X_1, X_2) = 1\{wX_1 + X_2 \geq 0\}$$

Threshold structure familiar from Bayesian decisions (Wald and Wolfowitz, 1948)

Key technical lemma:

- Explicit expression in $\tau^\star$ in the paper
- $\tau^\star$ is monotonic and continuous in c/s

$\Rightarrow$ technincal challenge: $\tau^\star$ depends on least-favorable distribution over $\Delta \sim \pi$. Problem is that least-favorable distribution changes with c.

Optimal decision: t-test structure

The following is the (almost sure) unique minimax optimal decision

$$\delta_1^\star(X_1) = \begin{cases} 1 & X_1/s \geq \tau^\star(c/s; w), \\ 0 & X_1/s \leq -\tau^\star(c/s; w), \\ \texttt{NA} & |X_1|/s < \tau^\star(c/s; w) \end{cases} \qquad \delta_2^\star(X_1, X_2) = 1\{wX_1 + X_2 \geq 0\}$$

Threshold structure familiar from Bayesian decisions (Wald and Wolfowitz, 1948)

Key technical lemma:

- Explicit expression in $\tau^\star$ in the paper
- $\tau^\star$ is monotonic and continuous in c/s

$\Rightarrow$ technincal challenge: $\tau^\star$ depends on least-favorable distribution over $\Delta \sim \pi$. Problem is that least-favorable distribution changes with c.

Q How to quantify "economic significance" without specifying c?

Introducing the Abstention Value (A-value)

The A-value is the smallest welfare cost at which an immediate decision becomes optimal:

$$A(X_1; s) = \inf\left\{c \geq 0 : \frac{|X_1|}{s} \geq \tau^\star(c/s)\right\}.$$

Introducing the Abstention Value (A-value)

The A-value is the smallest welfare cost at which an immediate decision becomes optimal:

$$A(X_1; s) = \inf\left\{ c \geq 0 : \frac{|X_1|}{s} \geq \tau^\star(c/s) \right\}.$$

Theorem

Every minimax-regret optimal rule satisfies, almost surely,

$$\delta_1^\star(X_1) = \texttt{NA} \quad \text{if and only if} \quad c < A(X_1; s).$$

Introducing the Abstention Value (A-value)

The A-value is the smallest welfare cost at which an immediate decision becomes optimal:

$$A(X_1; s) = \inf\left\{ c \geq 0 : \frac{|X_1|}{s} \geq \tau^\star(c/s) \right\}.$$

Theorem

Every minimax-regret optimal rule satisfies, almost surely,

$$\delta_1^\star(X_1) = \texttt{NA} \quad \text{if and only if} \quad c < A(X_1; s).$$

Properties

- $A(X_1; s) = sg(p)$, $p = 2(1 - \Phi(|X_1|/s))$ is the p-value, g is monotonic
- At a fixed standard error: a larger p-value gives a weakly larger A-value.
- At a fixed p-value: larger s (or larger $|X_1|$) implies larger A value

Remarks

- A-value is also a function of w
  - If future sample size unknown, set $w = 1$
  - If known, set $w = n_2/n_1$ with n_d sample size in each experiment d

Remarks

- A-value is also a function of w
  - If future sample size unknown, set $w = 1$
  - If known, set $w = n_2/n_1$ with n_d sample size in each experiment d
- Decision studied under pre-commitment as in Wald (1947).
  - Applies to funding and regulamentory agencies that pre-specified the decision function (e.g. Food and Admin, 2019; GiveWell, 2020)

Remarks

- A-value is also a function of w
  - If future sample size unknown, set $w = 1$
  - If known, set $w = n_2/n_1$ with n_d sample size in each experiment d
- Decision studied under pre-commitment as in Wald (1947).
  - Applies to funding and regulamentory agencies that pre-specified the decision function (e.g. Food and Admin, 2019; GiveWell, 2020)
- A-value depends on the unit of measurement.
  - Unit free version: divide by *outcome standard deviation* and interpret costs per standard deviation unit (in the paper)
  - Monetary measure: transform into monetary value

Examples

- **GiveWell**: philantropic funding agency $\Rightarrow$ We can use their internal estimates:
  - $\hat{b}$: estimated welfare return per dollar
  - κ: opportunity cost measured by GiveWell
  - K: implementation budget
  - C: experiment cost

Examples

- **GiveWell**: philantropic funding agency $\Rightarrow$ We can use their internal estimates:
  - $\hat{b}$: estimated welfare return per dollar
  - κ: opportunity cost measured by GiveWell
  - K: implementation budget
  - C: experiment cost

  $\Rightarrow$ It follows $X_1 = K(\hat{b} - \kappa), c = C\kappa \Rightarrow$ Experiment iff $c \leq A(X_1; s)$

Examples

- **GiveWell**: philantropic funding agency $\Rightarrow$ We can use their internal estimates:
  - $\hat{b}$: estimated welfare return per dollar
  - κ: opportunity cost measured by GiveWell
  - K: implementation budget
  - C: experiment cost

  $\Rightarrow$ It follows $X_1 = K(\hat{b} - \kappa), c = C\kappa \Rightarrow$ Experiment iff $c \leq A(X_1; s)$

- **FDA Approval**: when should we approve a given drug?
  - $\hat{b}$ estimated probability of survival
  - V value of statistical life
  - κ: cost per patient
  - N size of the market if drug approved
  - c: cost of additional trial

Examples

- **GiveWell**: philantropic funding agency $\Rightarrow$ We can use their internal estimates:
  - $\hat{b}$: estimated welfare return per dollar
  - κ: opportunity cost measured by GiveWell
  - K: implementation budget
  - C: experiment cost

  $\Rightarrow$ It follows $X_1 = K(\hat{b} - \kappa), c = C\kappa \Rightarrow$ Experiment iff $c \leq A(X_1; s)$

- **FDA Approval**: when should we approve a given drug?
  - $\hat{b}$ estimated probability of survival
  - V value of statistical life
  - κ: cost per patient
  - N size of the market if drug approved
  - c: cost of additional trial

  $\Rightarrow$ $X_1 = N(V\hat{b} - \kappa) \Rightarrow$ experiment iff $c \leq A(X_1; s)$

Outline

1. Setup
2. Allocating scarse resources
3. Empirical application
4. Conclusion

Outline

1. Setup
2. Allocating scarse resources
3. Empirical application
4. Conclusion

Which studies should receive follow-up?

A funding agency has J independent evaluations and capacity for q follow-ups. Study costs and welfare weights are equal [Unequal costs/weights in the paper] .

1. Put effects and standard errors in common welfare units.
2. Compute $A_j = A(X_{1j}; s_j)$ and rank them: $A_{(1)} \geq A_{(2)} \geq \cdots \geq A_{(J)}$.
3. Fund the q largest A-values. For the remaining, decide using $\text{sign}(X_{1j})$.

Which studies should receive follow-up?

A funding agency has J independent evaluations and capacity for q follow-ups. Study costs and welfare weights are equal [Unequal costs/weights in the paper] .

1. Put effects and standard errors in common welfare units.
2. Compute $A_j = A(X_{1j}; s_j)$ and rank them: $A_{(1)} \geq A_{(2)} \geq \cdots \geq A_{(J)}$.
3. Fund the q largest A-values. For the remaining, decide using $\text{sign}(X_{1j})$.

Near-optimal allocation

Let $\bar{s} = \max_j s_j$ the largest standard deviation. Then for $\boldsymbol{\delta}^{A,q}$ denoting the proposed rule

$$\sup_{\boldsymbol{\Delta}} \mathcal{L}(\boldsymbol{\delta}^{A,q}; \boldsymbol{\Delta}) - \inf_{\text{feasible } \delta} \sup_{\boldsymbol{\Delta}} \mathcal{L}(\boldsymbol{\delta}; \boldsymbol{\Delta}) \lesssim \bar{s} \min\left\{ \frac{1}{\sqrt{J}}, \frac{q}{J} \right\}$$

Which studies should receive follow-up?

A funding agency has J independent evaluations and capacity for q follow-ups. Study costs and welfare weights are equal [Unequal costs/weights in the paper] .

1. Put effects and standard errors in common welfare units.
2. Compute $A_j = A(X_{1j}; s_j)$ and rank them: $A_{(1)} \geq A_{(2)} \geq \cdots \geq A_{(J)}$.
3. Fund the q largest A-values. For the remaining, decide using $\text{sign}(X_{1j})$.

Near-optimal allocation

Let $\bar{s} = \max_j s_j$ the largest standard deviation. Then for $\boldsymbol{\delta}^{A,q}$ denoting the proposed rule

$$\sup_{\boldsymbol{\Delta}} \mathcal{L}(\boldsymbol{\delta}^{A,q}; \boldsymbol{\Delta}) - \inf_{\text{feasible } \delta} \sup_{\boldsymbol{\Delta}} \mathcal{L}(\boldsymbol{\delta}; \boldsymbol{\Delta}) \lesssim \bar{s} \min\left\{ \frac{1}{\sqrt{J}}, \frac{q}{J} \right\}$$

$$\mathcal{L}(\boldsymbol{\delta}^{A,q}; \boldsymbol{\Delta}) - \inf_{\text{infeasible } \delta} \sup_{\boldsymbol{\Delta}} \mathcal{L}(\boldsymbol{\delta}; \boldsymbol{\Delta}) \leq \mathbb{E}_{\boldsymbol{\Delta}}[A_{(q)}].$$

Remarks

- $\boldsymbol{\delta}^{A,q}$ achieves minimax rate in per-study precision s_j (in the paper)
- $\mathbb{E}[A_q]$ characterizes the "cost" of the constraint as a function of A-value
- For large J, we can construct a DGP so that p-value only ranking, standard error only ranking, Bayes optimal with $\Delta \sim \pi$ for common prior π have non-vanishing minimax regret.

Extensions

- Weigh symmetric prior $\Delta \sim \pi$ around zero against minimax solution:

$$A_{\lambda,\pi} = \inf\left\{c \geq 0 : \frac{|X_1|}{s_1} \geq \tau_\lambda(c)\right\},$$

$$\tau_\lambda(c) = \Phi^{-1}\Big((1-\lambda)\Phi(\underbrace{\tau^\star(c/s_1)}_{\text{minimax optimal}}) + \lambda\Phi(\underbrace{\tau_\pi(c; s_1, s_2)}_{\text{Bayes optimal}})\Big).$$

It controls minimax regret by factor $1-\lambda$ and Bayes risk by factor λ

Extensions

- Weigh symmetric prior $\Delta \sim \pi$ around zero against minimax solution:

$$A_{\lambda,\pi} = \inf\left\{c \geq 0 : \frac{|X_1|}{s_1} \geq \tau_\lambda(c)\right\},$$

$$\tau_\lambda(c) = \Phi^{-1}\Big((1-\lambda)\Phi(\underbrace{\tau^\star(c/s_1)}_{\text{minimax optimal}}) + \lambda\Phi(\underbrace{\tau_\pi(c; s_1, s_2)}_{\text{Bayes optimal}})\Big).$$

  It controls minimax regret by factor $1-\lambda$ and Bayes risk by factor λ

- Non-Gaussian X_1 (exponential distribution)
- More than one sequential wave
- Asymmetric regret function

Outline

1. Setup
2. Allocating scarse resources
3. Empirical application
4. Conclusion

Setup

- Study 40+ cash transfer programs across countries (Crosta et al., 2024)
- Measure the A-value and its monetary equivalent
- Assess strength of current evidence and what we would like to fund next

- Consider policy expansion for 10 million costs over 12 months
- Most of studies have A-value below 500,000 $
- Study settings where funding agencies can fund 10 additional studies

Cash transfers effect on consumption (42 studies)

For N households, a transfer of T dollars and $M = 12$ months of gains,

$$\text{break-even research cost} = NTM \cdot A, \qquad \text{Set : } NT = 10 \text{ million}$$

Cash transfers effect on consumption (42 studies)

For N households, a transfer of T dollars and $M = 12$ months of gains,

$$\text{break-even research cost} = NTM \cdot A, \qquad \text{Set : } NT = 10 \text{ million}$$

Administrative costs (24%)		
Statistic	A-value	Monetary equivalent ($)
Studies	42	—
Mean	0.230	276,364
Standard deviation	0.828	993,472
25th percentile	6.76×10^{-4}	811
Median	0.013	15,669
75th percentile	0.044	52,637

2010 USD PPP. The administrative-cost specification subtracts 24% over 12 months from each program's gross consumption estimate before computing A-values.

Example of countries selected: select 10 studies

Administrative costs (24%): illustrative programs

Country	$\lvert t_j \rvert$	s_j	Monetary A ($)	EB prior weight (%)	Top-ten selection			
					A	p	s_j	EB
Nepal	0.16	37.325	5,892,394	99.7	✓	✓	✓	–
Tanzania	0.21	2.900	402,328	65.9	✓	✓	–	✓
Bangladesh	0.90	10.575	304,698	96.2	✓	–	✓	–
Togo	1.14	4.250	71,042	80.6	✓	–	–	–

Example of countries selected: select 10 studies

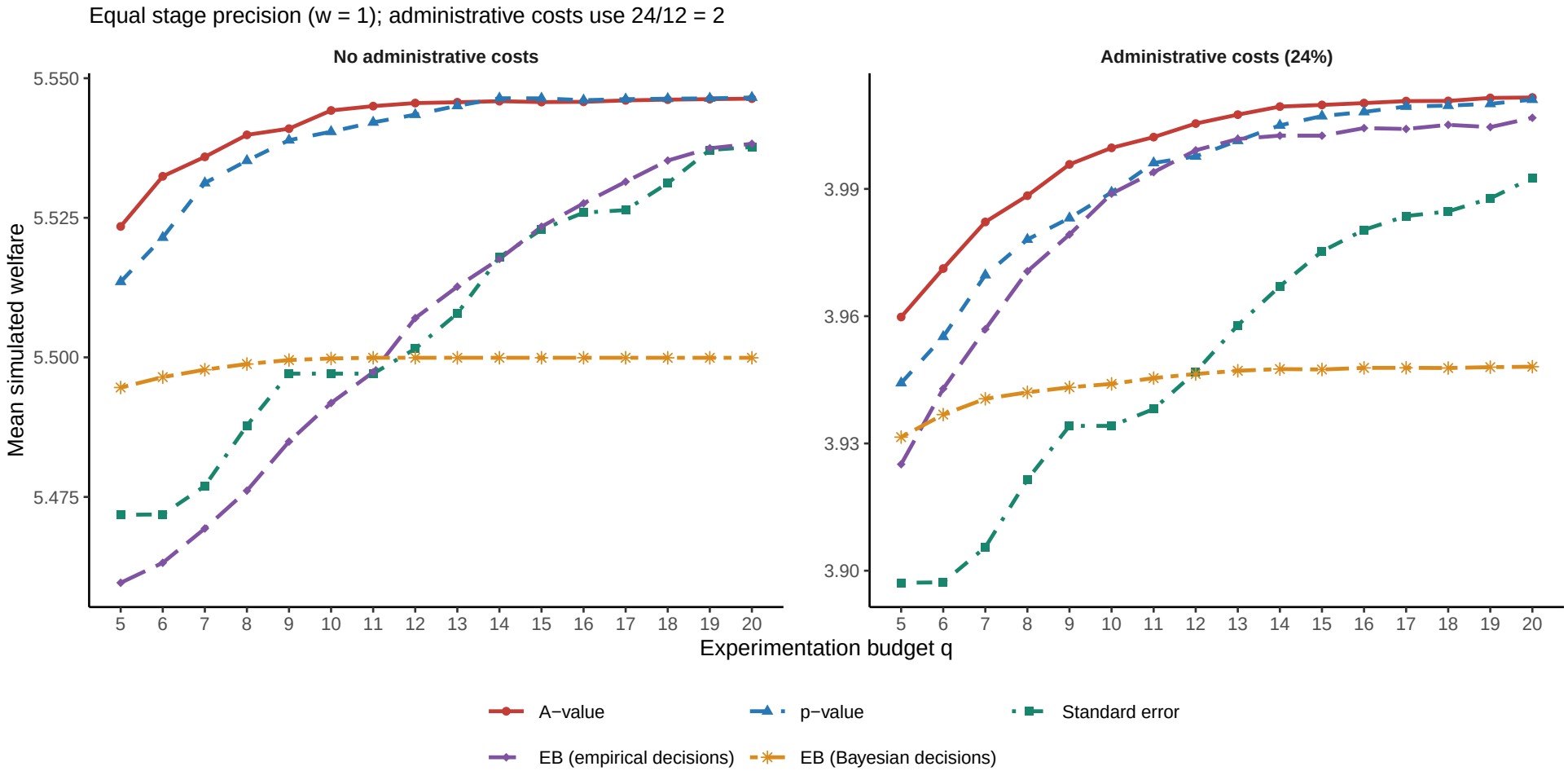
Administrative costs (24%): illustrative programs

Country	$\lvert t_j \rvert$	s_j	Monetary A ($)	EB prior weight (%)	Top-ten selection A	p	s_j	EB
Nepal	0.16	37.325	5,892,394	99.7	✓	✓	✓	–
Tanzania	0.21	2.900	402,328	65.9	✓	✓	–	✓
Bangladesh	0.90	10.575	304,698	96.2	✓	–	✓	–
Togo	1.14	4.250	71,042	80.6	✓	–	–	–
Liberia	1.23	3.900	51,984	77.7	–	–	–	✓
Rwanda	0.25	0.400	49,765	3.5	–	✓	–	✓
South Sudan	2.05	7.700	10,403	93.2	–	–	✓	–

Top-ten overlap with A (all 42 programs): p: 7/10 s_j: 6/10 EB: 3/10.

- p indicates p-value ranking, s_j indicates standard error ranking, EB indicates Empirical Bayes with Gaussian prior

Calibrated simulations



Generating A-value reports: recommended practice

Intervention: [name] *Welfare outcome:* [definition and units]

A. Evidence		**B. Calibration and costs**	
Estimate net of cost (standard error)	$[X_1\ (s_1)]$	Population affected	$[N]$
Per unit intervention cost	$[\kappa]$	Monetary value per unit	$[V]$
Two-sided p-value	$[p]$	Welfare horizon	$[H]$
Initial sample size (treatment probability)	$[n_1(\pi)]$	Intervention budget	$[K]$
Normalization	$[\sigma]$	Monetary A-value	$[C_A]$
Standardized A-value	$[A^\star]$	Original study cost	$[C_1]$
Context mean $A^\star$ (studies)	$[\overline{A}^\star\ (K)]$	Follow-up cost	$[C_{\text{exp}}]$

Generating A-value reports: recommended practice

Intervention: [name] *Welfare outcome:* [definition and units]

A. Evidence		**B. Calibration and costs**	
Estimate net of cost (standard error)	$[X_1\ (s_1)]$	Population affected	$[N]$
Per unit intervention cost	$[\kappa]$	Monetary value per unit	$[V]$
Two-sided p-value	$[p]$	Welfare horizon	$[H]$
Initial sample size (treatment probability)	$[n_1(\pi)]$	Intervention budget	$[K]$
Normalization	$[\sigma]$	Monetary A-value	$[C_A]$
Standardized A-value	$[A^\star]$	Original study cost	$[C_1]$
Context mean $A^\star$ (studies)	$[\overline{A}^\star\ (K)]$	Follow-up cost	$[C_{\text{exp}}]$

C. Sensitivity of the recommendation

Scenario	Standardized A-value	Monetary A-value	Follow-up cost	Recommend follow-up?
Baseline: $w = 1$	$[A^\star]$	$[C_A]$	$[C_{\text{exp}}]$	[Yes/No]
Alternative calibration: $[N, V, H]$	$[A^\star]$	$[C'_A]$	$[C_{\text{exp}}]$	[Yes/No]
Alternative design: $w = [\,\cdot\,]$	$[A_{\text{std}}(w)]$	$[C_A(w)]$	$[C_{\text{exp}}(w)]$	[Yes/No]

(Standardized A-value: $\frac{A}{\sqrt{n_1(\pi(1-\pi))}}$, n_1 experiment sample size, π treatment prob)

Outline

1. Setup
2. Allocating scarse resources
3. Empirical application
4. Conclusion

Conclusion: receipt for practice

- We study how to use data to learn what to learn next
- Goal is to measure strength of economic evidence
  - Think ahead about cost-benefit analysis involved
  - Compute A-value and its monetary equivalent
  - Compare with previous A-value findings (when available)
- Analysis useful both for researchers and funding agencies

Conclusion: receipt for practice

- We study how to use data to learn what to learn next
- Goal is to measure strength of economic evidence
  - Think ahead about cost-benefit analysis involved
  - Compute A-value and its monetary equivalent
  - Compare with previous A-value findings (when available)
- Analysis useful both for researchers and funding agencies

Applicability include

- Field experiments in development and funding anti-poverty programs
- Clinical trials/drug approval, health economics, public economics,...

Thank you!

References I

Abadie, A. (2020, June). Statistical nonsignificance in empirical economics. *American Economic Review: Insights 2*(2), 193–208.

Abadie, A., A. Agarwal, G. Imbens, S. Jia, J. McQueen, S. Stepaniants, and S. Torres (2023). Estimating the value of evidence-based decision making. *arXiv preprint arXiv:2306.13681*.

Adusumilli, K. (2026). How to sample and when to stop sampling: The generalised Wald problem and minimax policies. *The Review of Economic Studies 93*(1), 1–34.

Andrews, I. and J. Chen (2025). Certified decisions. *arXiv preprint arXiv:2502.17830*.

Arrow, K. J., D. Blackwell, and M. A. Girshick (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica 17*(3/4), 213–244.

Azevedo, E. M., A. Deng, J. L. Montiel Olea, J. Rao, and E. G. Weyl (2020). A/b testing with fat tails. *Journal of Political Economy 128*(12), 4614–000.

Azevedo, E. M., D. Mao, J. L. Montiel Olea, and A. Velez (2023). The A/B testing problem with gaussian priors. *Journal of Economic Theory 210*, 105646.

Bates, S., M. I. Jordan, M. Sklar, and J. A. Soloff (2026). Principal-agent hypothesis testing. *Journal of the American Statistical Association* (just-accepted), 1–38.

Bather, J. A. (1962). Bayes procedures for deciding the sign of a normal mean. *Mathematical Proceedings of the Cambridge Philosophical Society 58*(4), 599–620.

Benjamin, D. J., J. O. Berger, M. Johannesson, et al. (2018). Redefine statistical significance. *Nature Human Behaviour 2*(1), 6–10.

References II

Breza, E., A. G. Chandrasekhar, and D. Viviano (2025). Evidence aggregation with ignorance in mind: Learning what we do (not) know for archetypes discovery. Revised June 2026; revise and resubmit at Econometrica.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics 30*(3), 755–770.

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory 16*(1), 41–46.

Crosta, T., D. Karlan, F. Ong, J. Rüschenpöhler, and C. R. Udry (2024). Unconditional cash transfers: A bayesian meta-analysis of randomized evaluations in low and middle income countries. Technical report, National Bureau of Economic Research.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Food, U. and D. Admin (2019). Guidance document: Adaptive design clinical trials for drugs and biologics.

Fudenberg, D., P. Strack, and T. Strzalecki (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review 108*(12), 3651–3684.

GiveWell (2020, March). GiveWell's pre-analysis plan for the New Incentives RCT: Addendum on internal and external validity.

Grundy, P. M., M. J. R. Healy, and D. H. Rees (1956). Economic choice of the amount of experimentation. *Journal of the Royal Statistical Society: Series B (Methodological) 18*(1), 32–49. Discussion on pp. 49–55.

References III

Grundy, P. M., D. H. Rees, and M. J. R. Healy (1954). Decision between two alternatives—how many experiments? *Biometrics 10*(3), 317–323.

Hendren, N. and B. Sprung-Keyser (2020). A unified welfare analysis of government policies. *The Quarterly journal of economics 135*(3), 1209–1318.

Horowitz, J. L. and S. Lee (2025). Binary classification with the maximum score model and linear programming.

Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives 35*(3), 157–174.

Liang, A., X. Mu, and V. Syrgkanis (2022). Dynamically aggregating diverse information. *Econometrica 90*(1), 47–80.

Manski, C. F. (2021a). Econometrics for decision making: Building foundations sketched by Haavelmo and Wald. *Econometrica 89*(6), 2827–2853.

Manski, C. F. (2021b). Statistical inference for statistical decisions. *The American Statistician 75*(4), 447–456.

Manski, C. F. and A. Tetenov (2016). Sufficient trial size to inform clinical practice. *Proceedings of the National Academy of Sciences 113*(38), 10518–10523.

Maurice, R. J. (1957). A minimax procedure for choosing between two populations using sequential sampling. *Journal of the Royal Statistical Society: Series B (Methodological) 19*(2), 255–261.

References IV

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics 166*(1), 157–165.

Tetenov, A. (2016). An economic theory of statistical testing. CeMMAP Working Paper CWP50/16, Centre for Microdata Methods and Practice.

Viviano, D., K. Wüthrich, and P. Niehaus (2026). A model of multiple hypothesis testing. *The Review of Economic Studies*, rdag026.

Wald, A. (1947). *Sequential Analysis*. New York: John Wiley & Sons.

Wald, A. and J. Wolfowitz (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics 19*(3), 326–339.

Wald, A. and J. Wolfowitz (1950). Bayes solutions of sequential decision problems. *The Annals of Mathematical Statistics 21*(1), 82–99.