From Bayes' Billiard Table to Modern AI: Statistical Inference via Analytical and Simulation-Based Randomness

Minge Xie

Department of Statistics
Rutgers, The State University of New Jersey

The 13th Annual Joint Statistics Workshop
The ASA New Jersey Chapter and Bayer Clinical Statistics & Analytics
Bayer; October 9, 2026

Research supported in part by NSF Grants DMS-2311064, DMS-2319260, DMS-2515766, and NIH Grant R01GM157610.

Statistics is not Probability; rather, it is Data Science

Probability versus **Statistics** –

- **Probability** is a field of mathematics: Start with a model (and setup) and describe or predict random outcomes.
  - Example: Suppose we have a fair coin with a 50% chance of landing heads, and we toss it 10 times. How many times does it land on heads?

- **Statistics (Inference)** is data science: Start with observed outcomes (data) and work backward to infer (guess) the model.
  - Example: If we see 7 heads out of 10 tosses, is the coin fair?

In Bayes's time and after (∼150 years after his death), modern day "statistics" was called "inverse probability."

The Big Debate (18th Century) and Birth of Statistics

The Bayesian statistical method is named after **Thomas Bayes**, a Presbyterian minister, philosopher and mathematician

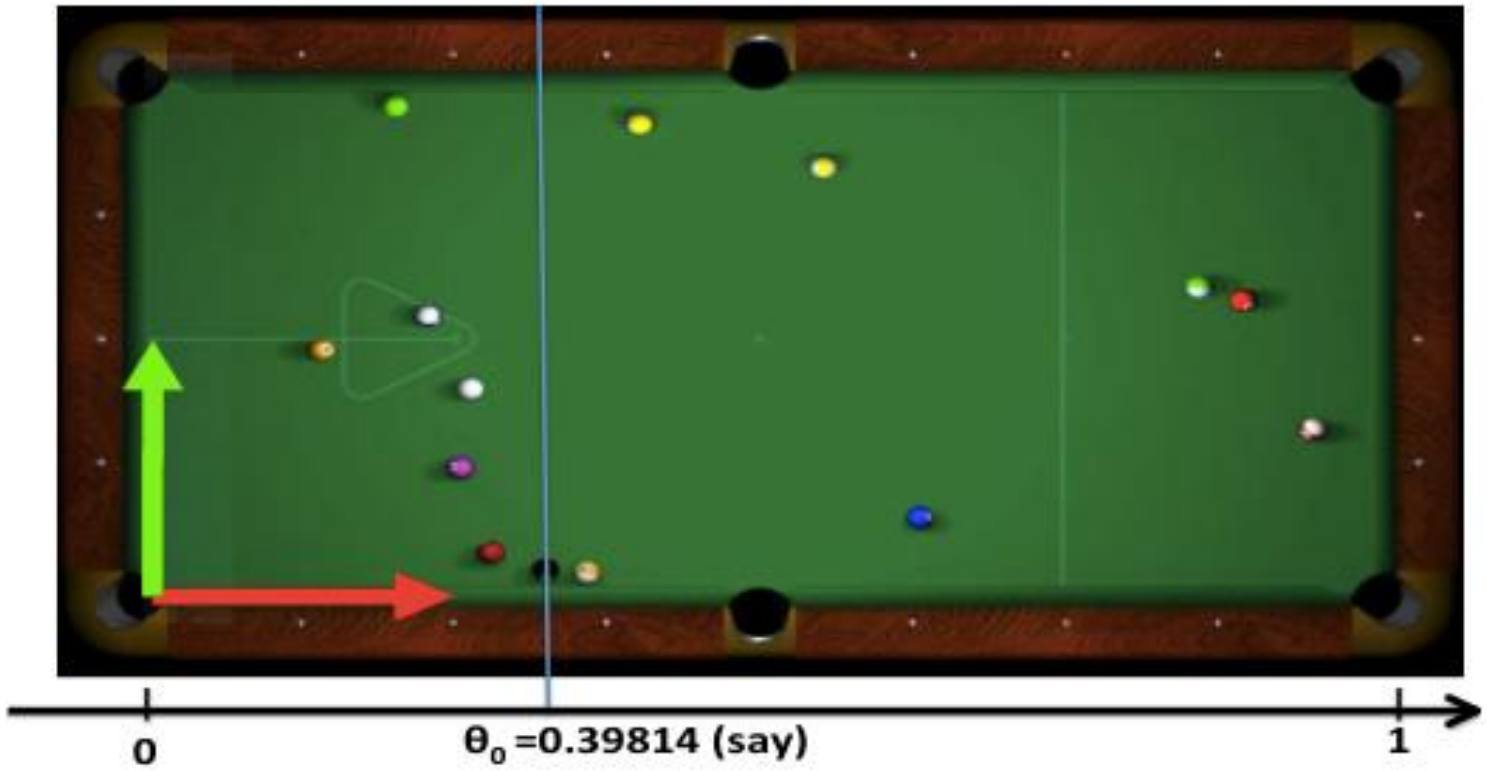
- ∼250 years ago in Britain and Europe, there was a debate among intellectuals (philosophers, priests, mathematicians, etc.):
  - Q: Based on what was observed in the world, can humans figure out what God's plan is?

- Thomas Bayes designed a **billiard ball experiment** to first answer the question – the answer is "YES"
  - Bayes completed but did not publish his seminal work, which was later discovered in his desk by his friend Richard Price after his death
    - (Bonus Q: Why didn't he publish it?)



Bayes' billiard-table model

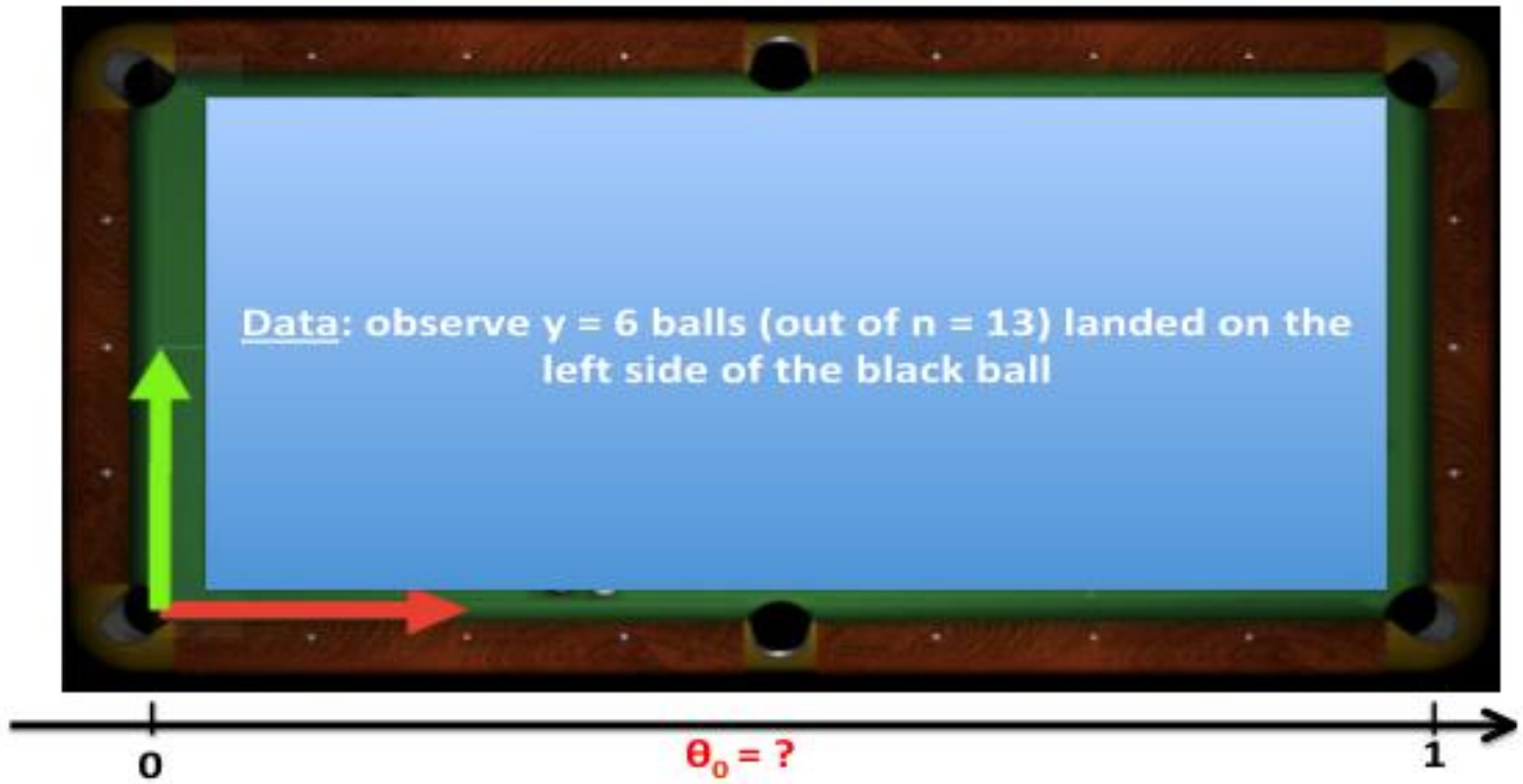
God's knowledge/full view



1) Randomly thrown one black ball onto a billiard table. Measure its landing place on the horizon (red direction) side of unit length $(0, 1)$.
2) Randomly thrown n additional balls onto the billiard table. Record how many (say y) of the n balls landed on the left side of the black ball. Take away the black ball.

Bayes' billiard-table model

Human's knowledge/observation



- **Sample data**: Take away the black ball ("truth") and only tell you that we observe $y = 6$ (out of $n = 13$) balls on the left side of the black ball
- **Goal**: Can you infer where the black ball was?

Bayes' billiard-table model

Bayes's billiard-table model In modern date mathematical expression

Model:

- The landing place $\theta \sim U(0,1)$
- The sample $y|\theta \sim Binomial(n,\theta)$

Observed (realized) data:

- The landing place θ_0 is a realization from $\theta \sim U(0,1)$
- The observed y_{obs} is a realization from $Binomial(n,\theta_0)$.

Goal:

- We are interested in inferring **what θ_0 is** (i.e., **what the true value $\theta_0 = 0.39814$ is** in the demo example).

Three (Two) Statistical Paradigms





Bayesian (175x - now) → Fisher's fiducial inference (193x - now, very small group) → Frequentist (194x - now, largest group)

- To a frequentist, unknown model parameter is fixed, the uncertainty of the parameter estimate is quantified by "repeated" experiments.
- Many Bayesians also think that the unknown parameter is a fixed value (a realization from a random mechanism). But since it is unknown and we are uncertain about it, so we can "think" (or treat) it as random to quantify the uncertainty. Thus we have distributions for the parameters of interest.

(Thornton and Xie, 2024, *Philosophy of Science*)

Three (Two) Statistical Paradigms





Bayesian (175x - now) → Fisher's fiducial inference (193x - now, very small group) → Frequentist (194x - now, largest group)

- To a frequentist, unknown model parameter is fixed, the uncertainty of the parameter estimate is quantified by "repeated" experiments.
- Many Bayesians also think that the unknown parameter is a fixed value (a realization from a random mechanism). But since it is unknown and we are uncertain about it, so we can "think" (or treat) it as random to quantify the uncertainty. Thus we have distributions for the parameters of interest.

(Thornton and Xie, 2024, *Philosophy of Science*)

Bayesian, Fiducial & Frequentist = BFF = Best Friends Forever

Why Statistics and Uncertainty Quantification Matter

Key scientific question: How can we make reliable scientific and clinical decisions when data are noisy, heterogeneous, limited, or incomplete?

Examples in drug development and clinical research:

- Clinical trials and real-world evidence
- Treatment effects and patient outcomes
- Biomarkers and high-dimensional molecular data
- AI and machine learning for prediction and decision support

Key decisions:

- Is a treatment effective and for whom?
- How precise and reliable is the estimated treatment effect?
- Which biomarkers or patient characteristics are truly informative?
- How should uncertainty affect clinical decisions?

Fundamental message:
Data alone are not enough — reliable scientific and clinical decisions require quantifying and communicating uncertainty.

An Evolving Landscape of Statistical Inference

Classical Statistical Inference: Analytic Approaches

- Probabilistic model (data-generating distribution) specification
- Likelihood-based inference
- Neyman–Pearson framework and decision theory

Emerging Direction: Inference from Synthetic (Artificial) Data

- Generative model specification (simulators, digital twins, etc.)
- Synthetic data as a surrogate to support principled inference and uncertainty quantification
- Inversion techniques for valid statistical inference and principled uncertainty assessment
- New inferential paradigms beyond the classical likelihood (e.g., likelihood-free inference)

Common Goal: Reliable scientific conclusions and decision-making under uncertainty.

Classical Statistical Modeling and Inference

$$\text{Data} \xrightarrow{\text{model assumption}} \text{Probability Model} \xrightarrow{\text{analytic expression}} \text{Likelihood} \xrightarrow{\text{large sample CLT}} \text{Inference}$$

Probability Models

- Parametric, semiparametric, and nonparametric models
- Regression, spatial, temporal, and hierarchical models

Inference

- Frequentist: estimation (e.g., MLE), testing, confidence intervals
- Bayesian: prior + likelihood $\rightarrow$ posterior
- Modern computation: MCMC, Monte Carlo, and Variational Inference

Goal: Reliable estimation, uncertainty quantification, prediction, and decision-making.

Example One: Johnson & Johnson's migraine headache study

The Annals of Applied Statistics
2013, Vol. 7, No. 1, 342–368
DOI: 10.1214/12-AOAS585
© Institute of Mathematical Statistics, 2013

INCORPORATING EXTERNAL INFORMATION IN ANALYSES OF CLINICAL TRIALS WITH BINARY OUTCOMES

BY MINGE XIE[1], REGINA Y. LIU[2], C. V. DAMARAJU AND WILLIAM H. OLSON

Rutgers, The State University of New Jersey, Rutgers, The State University of New Jersey, Ortho-McNeil Janssen Scientific Affairs LLC and Ortho-McNeil Janssen Scientific Affairs LLC

External information, such as prior information or expert opinions, can play an important role in the design, analysis and interpretation of clinical trials. However, little attention has been devoted thus far to incorporating external information in clinical trials with binary outcomes, perhaps due to the perception that binary outcomes can be treated as normally-distributed outcomes by using normal approximations. In this paper we show that these two types of clinical trials could behave differently, and that special care is needed for the analysis of clinical trials with binary outcomes. In particular, we first examine a simple but commonly used univariate Bayesian approach and observe a technical flaw. We then study the full Bayesian approach using

Example One: Johnson & Johnson's migraine headache study

- **Binomial clinical trial** on migraine headache
  - Randomized trial of two treatment groups:
    - n_1 subjects in A group (treatment/two drugs: topiramate & almotriptan);
    - n_0 subjects in B group (control/one drug: placebo & almotriptan)
  - Binary responses:

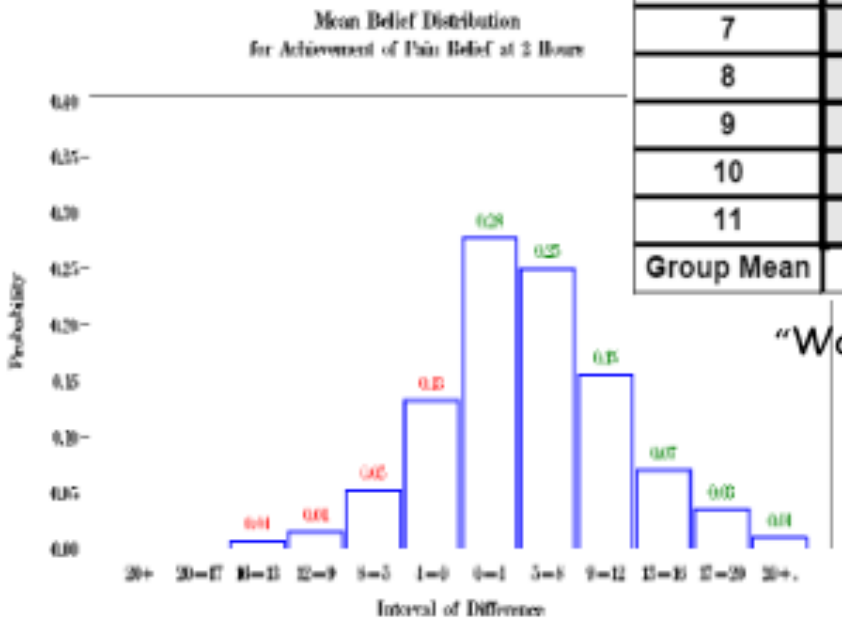
$$X_{1i} \sim Bernoulli(p_1) \quad \& \quad X_{0j} \sim Bernoulli(p_0)$$

- **Expert opinions** on the improvement $(\delta = p_1 - p_0)$
  - Being solicited from 11 experts, prior to the clinical trial.
  - Following the design of Parmar et al. (1994), Spiegelhalter et al. (1994)
- **Goal**: Incorporate expert opinions with information in the clinical data
  - Prior information (historical data) on p_0 may also be available

Example One: Johnson & Johnson's migraine headache study

Prior: Expert opinions on pain relief

Expert Opinions for Achievement of Pain Relief at 2 Hours												
	Worse (%)						Better (%)					
Investigator	20+	20-17	16-13	12-9	8-5	4-0	0-4	5-8	9-12	13-16	17-20	20+
1			5	5	10	30	30	15	5			
2				3	7	20	25	20	15	7	3	
3					10	15	20	20	20	10	5	
4			2	3	5	5	50	20	10	5		
5				5	10	15	15	30	20	5		
6						10	20	30	30	10		
7					5	20	50	20	5			
8						5	50	40	5			
9					5	5	20	30	20	10	10	
10					5	15	15	20	15	15	10	5
11						5	10	30	25	15	10	5
Group Mean			1	1	5	13	28	25	15	7	3	1



"Worse" ← Improvement $\delta = p_1 - p_0$ → "Betters"

(Courtesy of J&J)

Following Parmar et al. (1994) & Spiegelhalter et al. (1994)

Example One: Johnson & Johnson's migraine headache study

Bayesian approaches

- **Univariate on** $\delta = p_1 - p_0$ (Spiegelhalter et al., 1994 for normal clinical trials)
  - Normal prior: fit a normal curve to the histogram
  - Likelihood (estimated)

$$\hat{\ell}(\delta|\hat{\delta}) = -\frac{1}{2}\log(2\pi\hat{C}_\delta^2) - \frac{1}{2}(\delta - \hat{\delta})^2/\hat{C}_\delta^2$$

    where $\hat{\delta} = \bar{X}_1 - \bar{X}_0$ and $\hat{C}_\delta^2 = \bar{X}_1(1-\bar{X}_1)/n_1 + \bar{X}_0(1-\bar{X}_0)/n_0$.
  - Normal posterior (explicit expression)
- **Full Bayesian** — Joint model (p_0, p_1)
  - Bivariate beta prior (Olkin and Liu, 2003)
  - Likelihood of two independent binomial distributions
  - Posterior (not explicit expression): Use MCMC algorithm

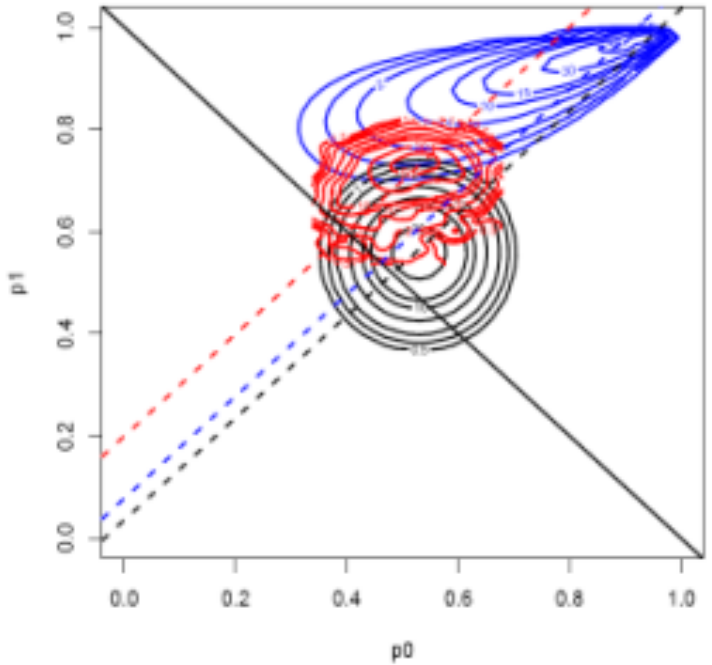
Example One: Johnson & Johnson's migraine headache study

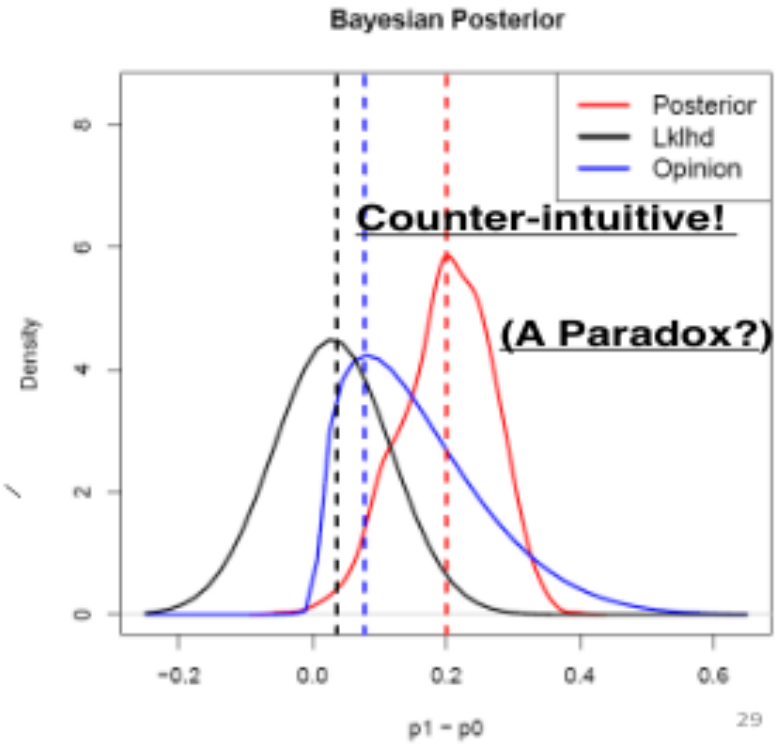
Issues in Bayesian approaches

- Strictly speaking, the **univariate approach** on $\delta = p_1 - p_0$ has a theoretical flaw – it is not supported by Bayesian theory!
  - Need to know $f(data|\delta)$ to apply Bayes formula, but this is NOT possible in the binomial trial
  - ⋆ Bayesian method is not good at "division of labor" (Efron 1986, Wasserman 2007)
- **Full Bayesian approach** jointly modeling (p_0, p_1) provides a comprehensive Bayesian solution
  - Theoretically sound and works in many situations
  - "Paradox" or discrepant posterior phenomenon: Marginal posterior of δ could be on the other side of both the data evidence and the marginal prior of δ!

Example One: Johnson & Johnson's migraine headache study

Bivariate Bayesian analysis (Full Bayesian approach)





Discrepant posterior phenomenon

Example One: Johnson & Johnson's migraine headache study

Summary on Bayesian methods

Empirical results:

- When histogram is close to bell shaped, all agree with each other.
- When histogram is skewed
  - Univariate Bayesian approach cannot capture the skewness
  - Full Bayesian approach produces a counterintuitive solution

Theoretical issues:

- A fundamental flaw exists in the univariate Bayesian approach
  - Not possible to find a conditional density function $f(data|\delta)$.
- A full Bayesian solution is to jointly model (p_0, p_1)
  - Is a theoretically sound approach.
  - (Bigger issue) Can produce a "counter-intuitive" discrepant/outlying posterior phenomenon.

In the era of modern data science and AI

Challenges: Classical method often relies on a correctly specified probabilistic model and tractable likelihood, but in modern data science and AI, data are increasingly complex and heterogeneous (e.g., images, voice, graphs, spatial–temporal data) and high-dimensional; models are often misspecified or have a large number of (unidentifiable) parameters, and the likelihood is either unavailable or computationally intractable due to this complexity and scale.

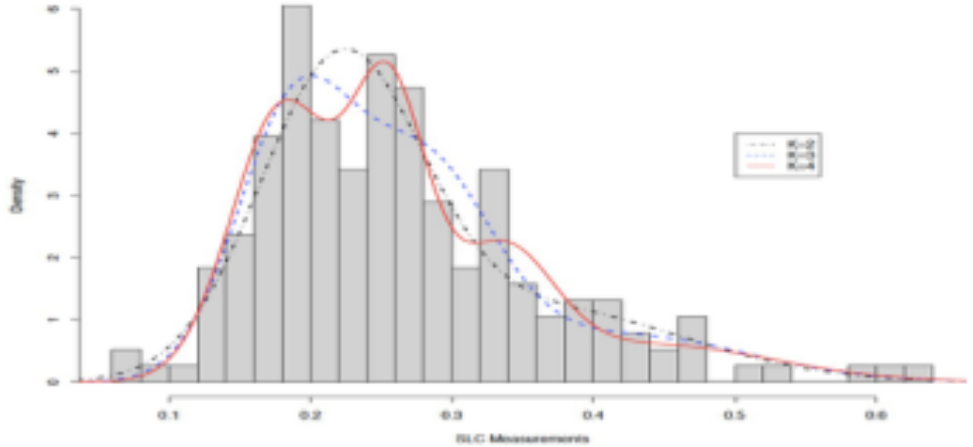
Why need new inference frameworks in DS/ML/AI era?

Big (biggest) challenge of statistics in DS/ML/AI era: To handle inference problems in which regularity conditions do not apply; i.e., those in which we cannot use large sample theories (e.g., CLT) or likelihood inference approaches, etc.? (Wasserman et al. '20)

- (Numerical) discrete or Non-numerical parameters: # of clusters; # of neighbors/nodes (in network); # of layers or nodes (in NN); Model A, B, ...; Tree partitions; etc.
- Non-numerical data: Graphical network; Image/voice/text data; etc.

Example A: Gaussian mixture model

Unknown # of components $\tau_0 = ?$



Histogram: observed numerical data

Example B: Graphical network model

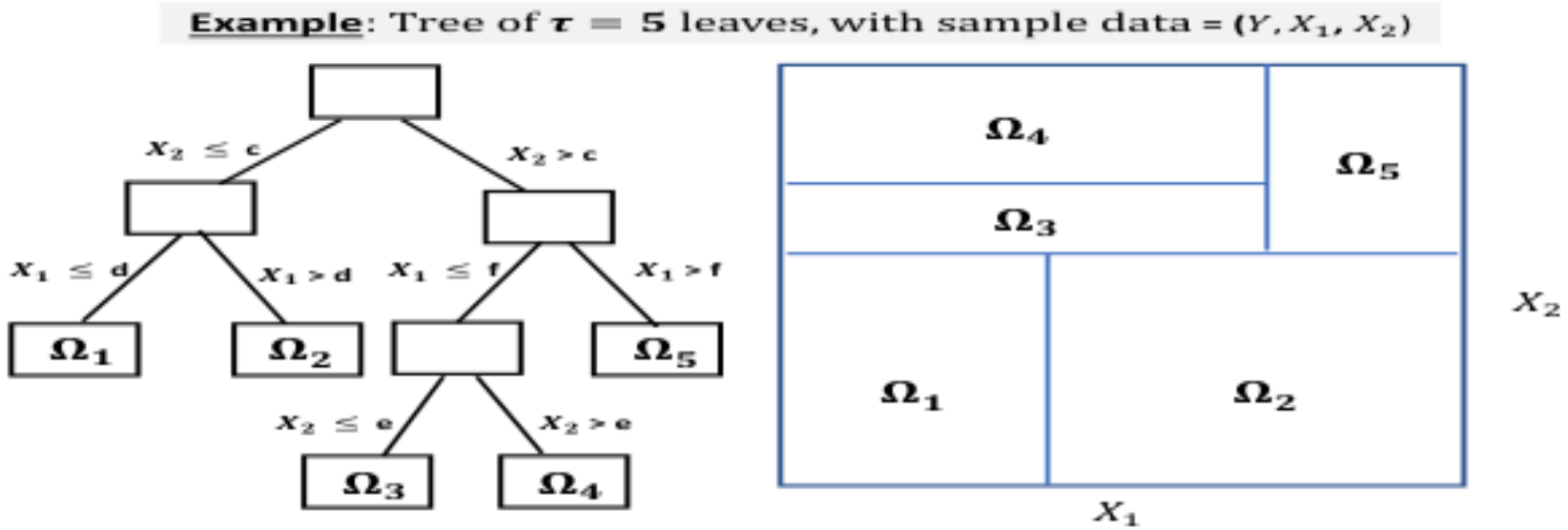
Unknown root node(s) $\tau_0 = ?$



(a) Observed network; (b) Root nodes revealed

Why need new inference frameworks in DS/ML/AI era?

Example C: Tree/recursive partition:



(Friedman et al. '09; Zhang & Singer '10)

- Sample data $\{(Y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ from a tree model

$$Y_i = \sum_{k=1}^{\tau_0} \mathbf{1}_{\{\mathbf{x}_i \in \Omega_k^{(0)}\}} (\mu_k^{(0)} + \sigma_k^{(0)} U_i), \quad U_i \sim N(0,1),$$

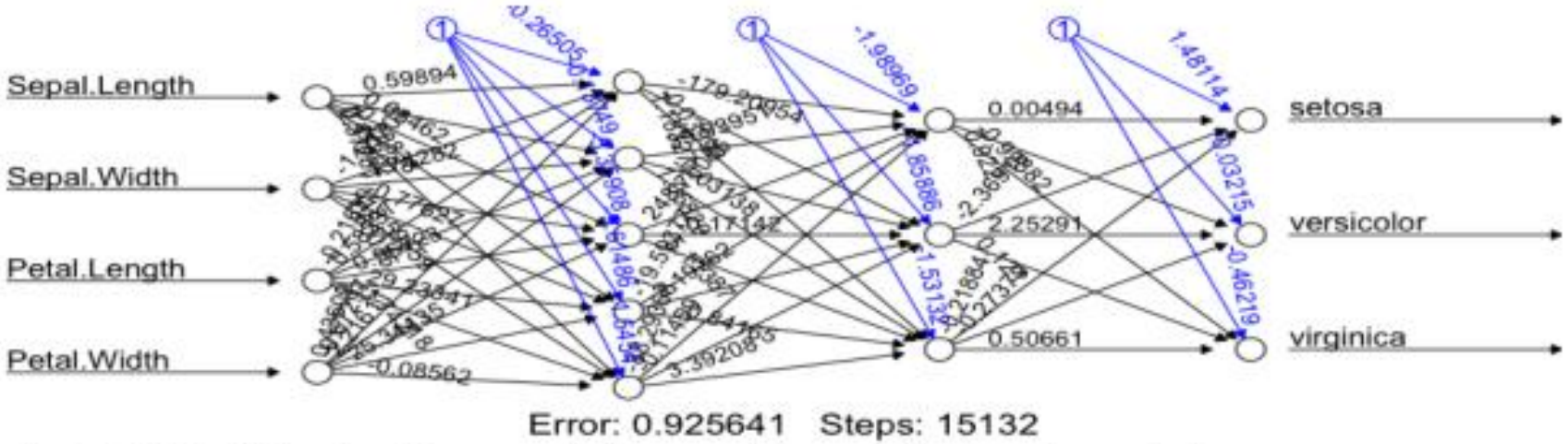
  where the model parameters are partition $B = \{\Omega_k : 1 \le k \le \tau\}$, $\mu = (\mu_1, \ldots, \mu_\tau)^T$, $\sigma = (\sigma_1, \ldots, \sigma_\tau)^T$ and $\theta = (B, \mu, \sigma)$; True value $\theta_0 = (B_0, \mu_0, \sigma_0)$.

  - Target: Find a confidence set for the 'true' tree $B_0 = \{\Omega_k^{(0)} : 1 \le k \le \tau_0\}$!

Principled random forest: Make prediction using ensemble trees in the conf set (far less number of trees and more interpretable!)

Why need new inference frameworks in DS/ML/AI era?

Example D: Neural Network Model:



- Sample data $\{(Y_i, X_i) : i = 1, \ldots, n\}$ from a Neural Network model

$$Y_i = f_\theta(X_i) + \epsilon_i = \mathbf{W}_{L+1}\mathbf{g}_{\gamma_L}(\mathbf{W}_L\mathbf{g}_{\gamma_{L-1}}(\ldots(\mathbf{W}_2\mathbf{g}_{\gamma_1}(\mathbf{W}_1 X_i)))) + \epsilon_i, \quad \epsilon_i \sim N(0,1),$$

where the model parameters are

$$\boldsymbol{\theta} = \big(L, (K_1, \ldots, K_L), (\mathbf{W}_1, \ldots, \mathbf{W}_L), (\gamma_1, \ldots, \gamma_L)\big),$$

including number of layer L, numbers of vertices $(K_1, \ldots, K_L)$, weights $(\mathbf{W}_1, \ldots, \mathbf{W}_L)$ and shifts $(\gamma_1, \ldots, \gamma_L)$. (Here, $\mathbf{W}_j$ is a $K_j \times K_{j-1}$ matrix; γ_j is a $K_j \times 1$ vector)

- Target: Find a confidence set for 'architecture' parameter $\tau_0 = (L^{(0)}, K_1^{(0)}, \ldots, K_L^{(0)})$!

Why need new inference frameworks in DS/ML/AI era?

Example E: High-dimensional linear regression (all inference problems, including model selection uncertainty): Response $\mathbf{Y} \in \mathbb{R}$ and covariates $X^{p\times 1}$; $p >> n$, $n =$ sample size

$$\mathbf{Y} = X^\top \beta_0^{full} + \sigma_0 U = X_{\tau_0}^\top \beta_0 + \sigma_0 U, \quad U \sim N(0,1), \quad \text{(conventional model)}$$

- $\tau_0 =$ true (sparse) model; $\beta_0 = \beta_{\tau_0}$; $p =$ dimension of β_0^{full}

Unaddressed questions: 1) Quantify estimating uncertainty of τ_0; 2) Finite sample inference ($n \not\to \infty$)

What we can offer:

- Level $1-\alpha$ confidence set for the true (sparse) model τ_0
- Level $1-\alpha$ confidence set for β_0^{sub} (any scalar, sub-vector or the full vector of β_0^{full})
- Joint level $1-\alpha$ confidence set for the mixed type of parameters (τ_0, β_0).
- Performance guaranteed finite & large sample supporting theory; No "min-signal" requirements

Relevant discussions:

- Address the question (side-step the topic) of poster selection inference
- Empirical performance better than state-of-the-art debiased methods especially for $\beta_{0,j} \neq 0$.

Why need new inference frameworks in DS/ML/AI era?

Example F: High dimensional binary regression (classification) model of unknown form
True Model: Binary response $\mathbf{Y} \in \{\mathbf{0}, \mathbf{1}\}$ and covariates $X^{p\times 1}$; $p >> n$, $n =$ sample size

$$P(y = 1|X) = \mu(X) \quad \text{for an unknown/unstructured } \mu(\cdot)$$

Working model: Due to lack of info on $\mu(\cdot)$, we fit a sparse regression model regardless!

$$P(y = 1|X) = g(X_\tau^\top \beta_\tau)$$

where $g(\cdot) =$ a given link function (logistic, probit, etc.) of a conventional model

- Define oracle model and oracle parameters (τ_0, β_{τ_0}): 'Closest' $g(X_\tau^\top \beta_\tau)$ to $\mu(x)$!

What we can offer:

- Confidence sets for τ_0, any linear function of β_{τ_0}, $g(X_{new,\tau_0}^\top \beta_{\tau_0})$, and $\mathbf{1}_{\{\mu_0(X_{new})>c\}}$.
- If $\mu_0(X) \equiv g(X_{\tau_0}^\top \beta_{\tau_0})$, we can recover the true sparsity τ_0 and make inference for τ_0, any linear functions of β_{τ_0}, and $\mu_0(X_{new})$
  (Similar results can be obtained, except under slightly stronger conditions)

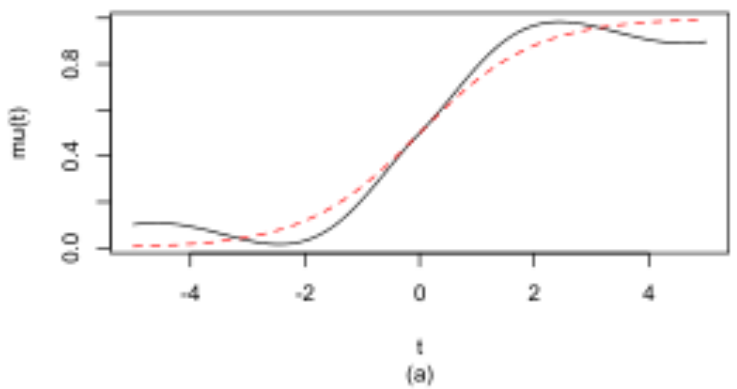
Why need new inference frameworks in DS/ML/AI era?

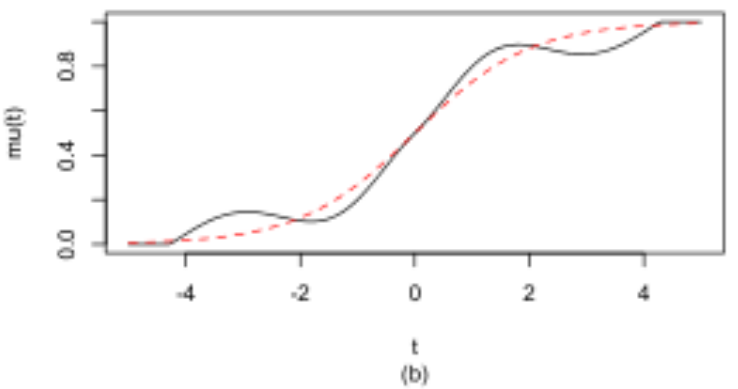
Example F: High dimensional binary regression (classification) model of unknown form
(continue ...)

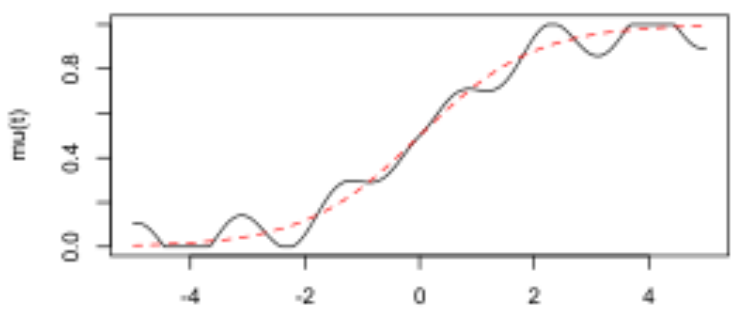
A numerical study example that we considered at the beginning:

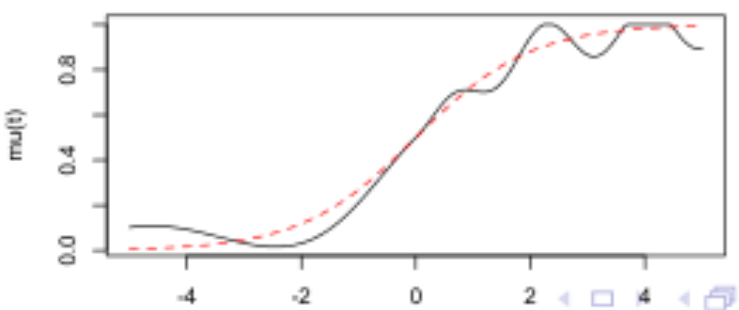
$$\mu(X) = \max\left\{0, \min\left\{1, g^{-1}(X^\top\gamma) + 0.2|g^{-1}(X^\top\gamma) - \frac{1}{2}|\sin(X^\top\omega)\right\}\right\}.$$

where $g(t) = \log\frac{t}{1-t}$.









Why need new inference frameworks in DS/ML/AI era?

Real Data Analysis: Single-cell RNA-seq Data (Shalek et al. 2014)

A subset of single-cell RNA-seq data consisting of:

- 96 cells stimulated by the pathogenic component PIC (a viral-like double-stranded RNA);
- 96 control cells without stimulation;
- Gene expression measured six hours after stimulation.

Data used: Further cleaned version from Cai *et al.* (2021), with

$$n = 192 \text{ samples}, \quad p = 679 \text{ genes}.$$

Model: Binary regression of Y on $\mathbf{X}$ with an unknown and unstructured form:

$$\mathbb{P}(Y = 1 \mid \mathbf{X}) = \mu(\mathbf{X}).$$

(The class of working models: Sparse logistic regression $\text{logit}\{P(y = 1|X)\} = X_\tau^\top \beta_\tau$)

Why need new inference frameworks in DS/ML/AI era?

Real Data Analysis: Single-cell RNA-seq data

- Oracle model set (seven models)

Genes	Selected Models						
RSAD2	•	•	•	•	•	•	•
AK217941	•	•	•	•	•	•	•
IFIT1					•	•	
BC044745			•	•	•	•	•
ACTB		•	•			•	•
HMGN2							•

Five out of the six (except AK217941) were studied and known to be associated with immune systems in the immunological literature

- Single gene (marginal β_j) analysis:
  - Our method identified two genes RSAD2 and AK217941 as significant
  - Debiased method only identified single gene RSAD2 as significant (Cai et al 2021)

Why need new inference frameworks in DS/ML/AI era?

Why need new framework?

For these irregular (discrete/non-numerical) types of parameters and/or irregular (non-numerical) types of data?

- Point estimator (if exists): do not have uncertainty quantification (CLT does not apply)
- Bootstrap methods: do not work (CLT does not apply)
- Bayesian methods (credible sets): highly sensitive to prior choices; no frequency coverage results (even with $n \to \infty$, BvM/CLT does not apply)

Emerging new frameworks: simulation and artificial-sample inspired/based inference methods

- General: Apply to all settings, including regular inference problems where the model parameter is an inner point of a continuous parameter space
- Particularly effective for making inference for those non-trivial (irregular) inference problems

Additional attempt: "All models are wrong but some are useful." (G.E.P. Box) — How can statistical inference still be relevant for new data DS/ML/AI problems?

Key Intuition: Using Artificial Samples for Uncertainty Quantification and Statistical Inference

A New Inferential Perspective

Core idea:

- Generate many artificial datasets that mimic the data-generating mechanism underlying the observed sample.
- Treat these artificial datasets as reference worlds for quantifying uncertainty.
- Draw inference by comparing the observed dataset with a collection of plausible artificial counterparts.

Underlying matching principle

$$\mathcal{Y}^{sim} \overset{dist}{\approx} \mathcal{Y}_{obs} \quad (\text{or} \quad \theta(\mathcal{Y}^{sim}) \overset{dist}{\approx} \theta(\mathcal{Y}_{obs}))$$

Key Intuition: Using Artificial Samples for Uncertainty Quantification and Statistical Inference

A New Inferential Perspective

Core idea:

- Generate many artificial datasets that mimic the data-generating mechanism underlying the observed sample.
- Treat these artificial datasets as reference worlds for quantifying uncertainty.
- Draw inference by comparing the observed dataset with a collection of plausible artificial counterparts.

Underlying matching principle

$$\mathcal{Y}^{sim} \overset{dist}{\approx} \mathcal{Y}_{obs} \quad (\text{or} \quad \theta(\mathcal{Y}^{sim}) \overset{dist}{\approx} \theta(\mathcal{Y}_{obs}))$$

Related (existing) developments in statistics — not from nowhere

- **Artificial-sample-based methods:** Bootstrap, Approximate Bayesian Computation (ABC), Knockoffs, etc.
- **Inversion-based methods:** Conformal Prediction, Generalized Fiducial Inference (GFI), etc.

Approximate Bayesian Computation (ABC)

Motivation: Likelihood function is unavailable or computationally intractable (e.g., in complex differential equation models).

Bayesian model setup with a prior $\pi(\theta)$ and a sample distribution P_θ:

$$\theta \sim \pi(\theta), \qquad \mathcal{Y}|\theta \sim P_\theta.$$

ABC uses simulation to "obtains" a posterior distribution (without using Bayes formula!)

Algorithm:

1. Draw $\theta^b \sim \pi(\theta)$, then simulate $\mathcal{Y}^b|\theta^b \sim P_{\theta^b}$,
2. Accept θ^b if
$$d(S(\mathcal{Y}^b), S(\mathcal{Y}_{obs})) \leq \varepsilon.$$
3. Repeat 1 & 2

where:

- $S(\cdot)$ = summary statistic,
- $d(\cdot, \cdot)$ = discrepancy measure,
- ε = tolerance threshold.

Why ABC Works

Core principle: Parameters that generate artificial data resembling the observed sample should receive high posterior support.

If:

- summaries are informative (i.e., sufficient statistics),
- tolerance ε is very small,

then accepted simulations approximately satisfy

$$\mathcal{Y}^{sim} \approx \mathcal{Y}_{obs}.$$

Hence the accepted parameters θ^b approximate draws from

$$\pi(\theta \mid \mathcal{Y}_{obs}).$$

In the idealized limit:

$$\varepsilon \to 0,$$

and with $S(Y)$ being a sufficient statistic,

$$\pi_{ABC}(\theta \mid \mathcal{Y}_{obs}) \to \pi(\theta \mid \mathcal{Y}_{obs}).$$

Interpretation: ABC uses artificial randomness generated from the model to reproduce the stochastic structure of the observed data.

Two New Artificial-Sample- or Simulation-Based Inference Methods

Both inspired by the ABC method – both frequentist and sidestep ABC issues (i.e., sufficient statistics, choice of ϵ, low computing efficiency)

- **Repro samples method**

  One-sentence explanation (confidence sets): *For a candidate parameter value, generate many artificial datasets under that value. If the observed dataset appears sufficiently consistent (or conformal) with these artificial datasets, retain the parameter value; otherwise, exclude it.*

- **Simulation Based Inference (SBI)**

  Algorithm:

  Step 1: Simulate artificial parameter & data pairs $(\theta_j, \mathcal{Y}_j)$ *for* $\theta_j \sim \pi(\theta)$, $\mathcal{Y}_j \mid \theta_j \sim P_{\theta_j}$, $j = 1, \ldots, N$

  Step 2: Construct a classification dataset $\{(Z_j, \mathcal{Y}_j, \theta_j)\}_{j=1}^{2N}$, *where* $Z_j = 1$ *for matched pairs* $(\mathcal{Y}_j, \theta_j)$ *and* $Z_j = 0$, *for mismatched pair* $(\mathcal{Y}_j, \theta_j')$, $\theta_j' \neq \theta_j$.

  Step 3: Run a ML classifier to classify $Z \sim (\mathcal{Y}, \theta)$ *to classify whether* $\mathcal{Y}$ *is paired with its true* θ

  Outcome: Estimated (trained) classifier function $\widehat{h}(\mathcal{Y}, \theta) = \widehat{\mathbb{P}(Z = 1 \mid \mathcal{Y}, \theta)}$ are used to obtain (estimate) the likelihood or ABC posterior

$$\widehat{p}_\theta(\mathcal{Y}_{obs}) \propto \frac{\widehat{h}(\mathcal{Y}_{obs}, \theta)}{1 - \widehat{h}(\mathcal{Y}_{obs}, \theta)} \quad \text{or} \quad \pi_{ABC}(\theta|\mathcal{Y}_{obs}) \propto \pi(\theta) p_\theta(\mathcal{Y}_{obs})$$

General Setup: Generative Model & Inference goal

Random sample $Y \in \mathcal{Y}$; Observed data y_{obs}; Any types of parameters θ

- **Generative Model**: (G a given general function)

$$Y = G(\theta, U)$$

where an unknown random "error term" $U \sim D(\cdot)$ (known distribution).

  - Realized version (unobserved realization u^{rel} & unknown β_0):

$$y_{obs} = G(\theta_0, u^{rel})$$

Minor Assumption†: Given θ, we can generate the corresponding Y

Examples: (a) **Regular (statistics textbook) setup**: $Y \sim f_\theta(t) \Rightarrow Y = G(\theta, U)$, where $G(\theta, U) = F_\theta^{-1}(U)$, $U \sim U(0,1)$, $F_\theta(y) = \int_{\infty}^{y} f_\theta(t)dt$

(b) **High-dimensional regression**: $Y = X_\tau^\top \beta_\tau + \sigma U \Rightarrow Y = G(\theta, U)$; where $G(\theta, u) = X_\tau^\top \beta_\tau + \sigma u$, $\theta = (\tau, \beta_\tau^\top, \sigma)^\top$ & $U \sim N(0, I)$

(c) **Others; e.g., differential equation model, graphical models, tree model, pre-trained AI generative model (LLMs, DNNs, ...)**

- **Inference goal**: Get a guaranteed level $1 - \alpha$ confidence set for β_0 or a function $h(\beta_0)$

(†Possible extension to even misspecified models)

More Generally: Algorithmic or Estimation Equation Models

Note that:

$$y_{obs} = G(\theta_0, u^{rel}) \quad \stackrel{implies}{\Longrightarrow} \quad \theta_0 = \arg\min_{\theta} \|y_{obs} - G(\theta, u^{rel})\|$$

or more generally

$$\theta_0 = \arg\min_{\theta} L\left(y_{obs}, G(\theta, u^{rel})\right)$$

for a continuous loss function $L(y, y')$.

More Generally: Algorithmic or Estimation Equation Models

Note that:

$$y_{obs} = G(\theta_0, u^{rel}) \quad \stackrel{implies}{\Longrightarrow} \quad \theta_0 = \arg\min_{\theta} \|y_{obs} - G(\theta, u^{rel})\|$$

or more generally

$$\theta_0 = \arg\min_{\theta} L\big(y_{obs}, G(\theta, u^{rel})\big)$$

for a continuous loss function $L(y, y')$.

- All developments remain intact for sample data from
  - an Algorithmic Model: i.e., the model parameter θ can be expressed as a function of Y and U:

    $$\theta = A(Y, U)$$

    Example: $A(Y, U) = \arg\min_{\theta} \|y_{obs} - G(\theta, u^{rel})\|$
  - an Estimating Equation Model: i.e., the triplet (Y, θ, U) satisfies:

    $$H(Y, \theta, U) = 0$$

    Example: $\sum_{i=1}^{n} \mathbf{1}(Y_i < \theta) = U$ with unknown F, for $Y_i \sim F(\cdot)$, qth quantile $\theta = F^{-1}(q)$, and $U \sim Bin(n, q)$

  for a given algorithm A or a function H

Example Two: Change-points Detection in Data Science

- **Change-points**: time points at which the data-generating behavior changes abruptly.
- **Classical focus**: much of the literature focuses on *detecting* and *localizing* change-points, producing point estimates of the number and locations of structural breaks.
- **Uncertainty question**: after obtaining estimated change-points, how confident are we about

how many changes occurred, and where they occurred?

- **Our goal**: provide uncertainty quantification for both the number and the locations of multiple change-points through confidence sets.

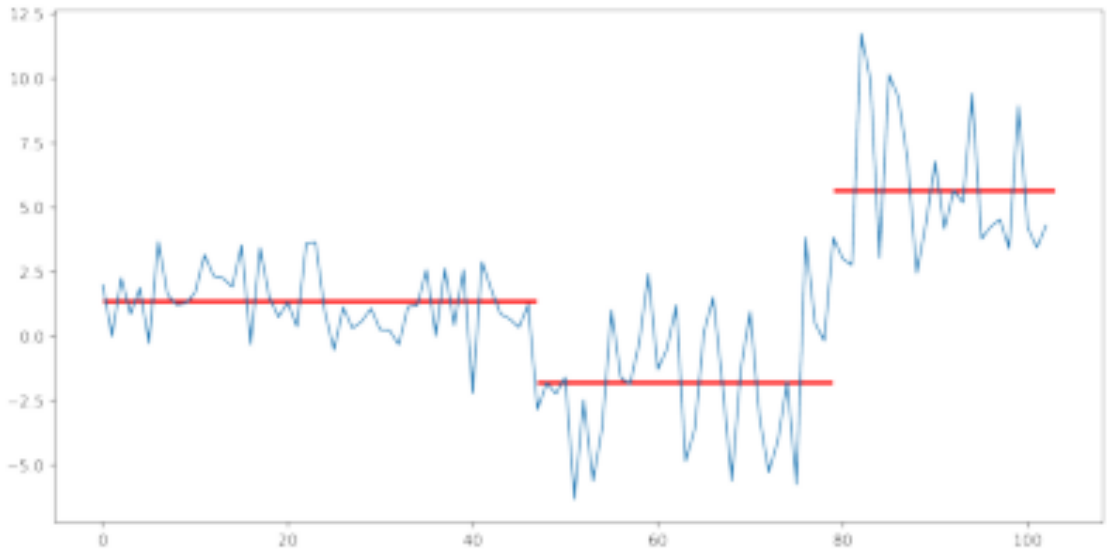
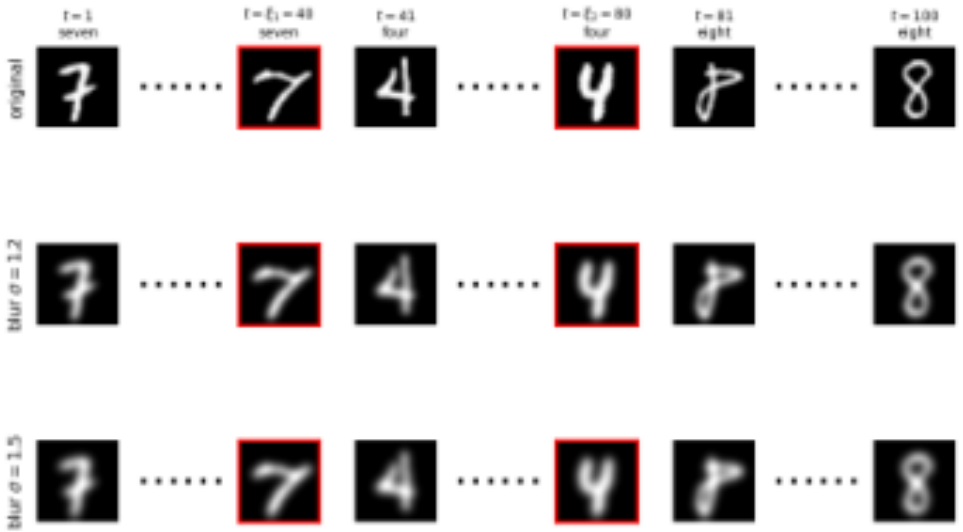


Figure: US ex-post real interest rate

Example Two: Change-points Detection in Data Science

Detection and uncertainty quantification in change-point detection with unknown number of change-points – Sequential image streams present major challenges for change-point inference because of High-dimensional observations, No tractable likelihood, etc.

MNIST Sequential Digits



CIFAR Sequential Animals



MNIST Confidence Regions

MNIST	CP1	CP2
Original	(40)	(80)
Slightly blurred	(39–42)	(79–80)
Moderately blurred	(28–42)	(68–81)

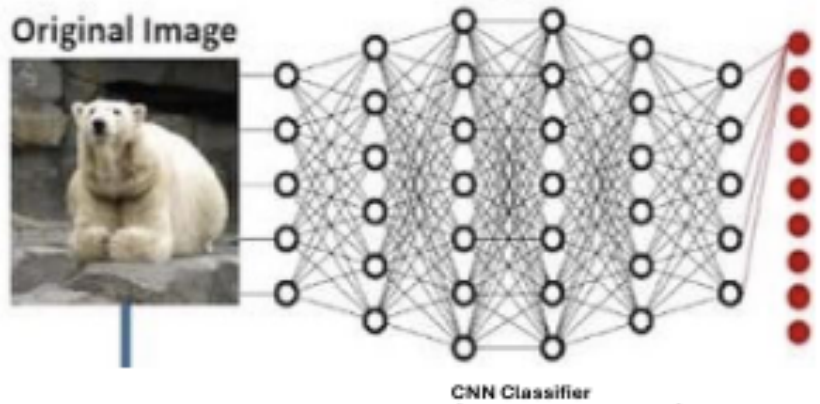
CIFAR Confidence Regions

CIFAR	CP1	CP2
Original	(39–40)	(80–81)
Slightly blurred	(30–50)	(69–90)
Moderately blurred	(31–52)	(66–90)

Uncertainty expands adaptively as signal quality deteriorates.

Example Two: Change-Point Detection in Data Science

Analysis pipeline for image change-point detection and inference:



- *CNN representation of images.* Pass each image through a CNN classifier and use the resulting softmax probability vector $\mathbf{x}_t$ as the working observation.
- *Development of a working generative model.* Conditional on the change-point locations, observations within each segment are modeled using an unknown distribution F_j, estimated via the ECDF (bootstrap).
- *Valid statistical inference.* Apply Repro Samples/SBI methods to infer both the unknown number of change-points and their locations.

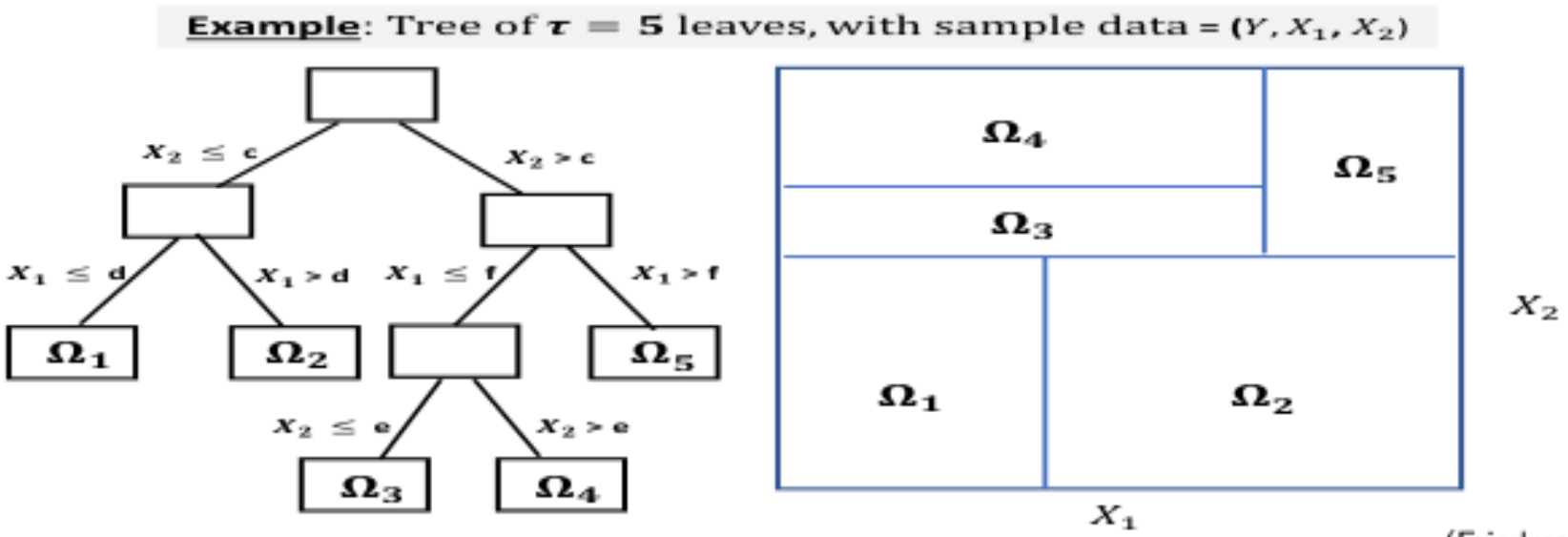
Takeaways and discussion:

- "Machine learning $\oplus$ repro-sample methodology" provides an effective framework for change-point detection and inference in all types of data, including image streams.
- Computationally feasible: the overall cost, including CNN processing and repro-sample inference, is comparable to that of standard bootstrap procedures.

(*Research supported by the U.S. NSF*)

Example Three: Principled Ensemble Forest

Example C: Tree/recursive partition:



(Friedman et al. '09; Zhang & Singer '10)

- Sample data $\{(Y_i, x_i) : i = 1, \ldots, n\}$ from a tree model

$$Y_i = \sum_{k=1}^{\tau_0} \mathbf{1}_{\{x_i \in \Omega_k^{(0)}\}} (\mu_k^{(0)} + \sigma_k^{(0)} U_i), \quad U_i \sim N(0,1),$$

where the model parameters are partition $B = \{\Omega_k : 1 \le k \le \tau\}$, $\mu = (\mu_1, \ldots, \mu_\tau)^T$, $\sigma = (\sigma_1, \ldots, \sigma_\tau)^T$ and $\theta = (B, \mu, \sigma)$; True value $\theta_0 = (B_0, \mu_0, \sigma_0)$.

- Target: Find a confidence set for the 'true' tree $B_0 = \{\Omega_k^{(0)} : 1 \le k \le \tau_0\}$!

Principled random forest: Make prediction using ensemble trees in the conf set

(Far less number of trees and more interpretable!)

Example Three: Principled Ensemble Forest

Principled random forest (PRF) vs random forest (RF)

Three-Leaf Tree Model: Data $\{Y_i, x_i\}_{i=1}^n$

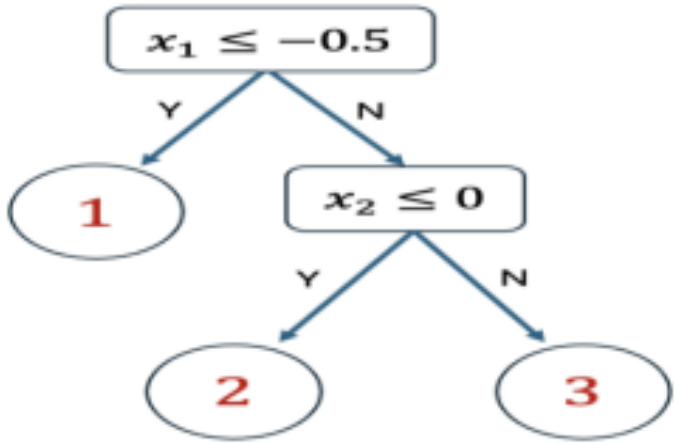
$Y_i = \sum_{k=1}^3 \mu_k \mathbf{1}\{g_i = k\} + U_i,\ U_i \overset{iid}{\sim} N(0,1)$

where $\mu = (\mu_1, \mu_2, \mu_3) = (2, 0.2, -0.2)$

$g_i = g(x_i)$ = membership assignment of i

Inference target: $\mu_{new} = E(Y_{new}|x_{new})$,

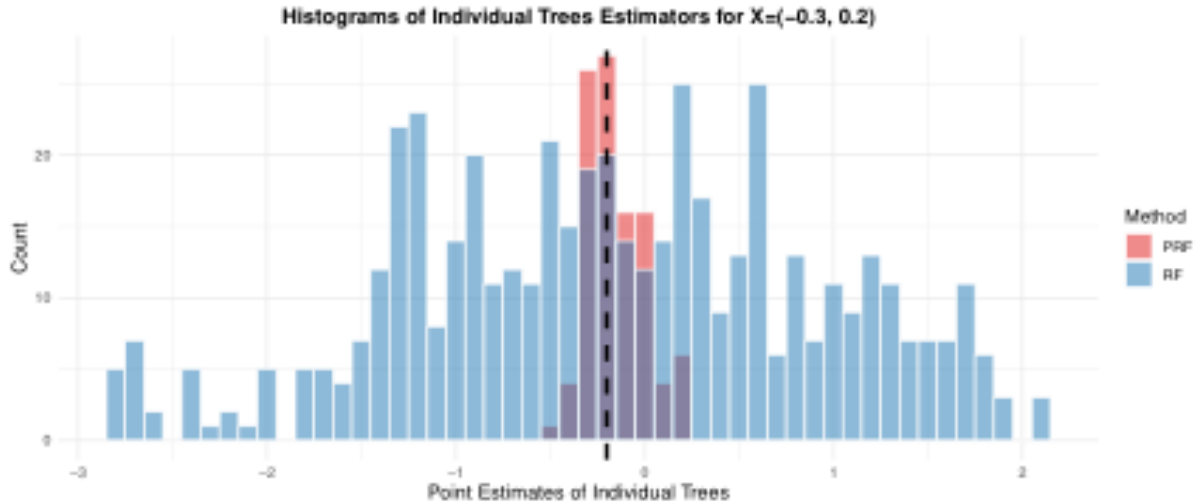
(e.g., $x_{new} = (-0.3, 0.2)$, $\mu_{new} = -0.2$)



Numerical results (only point estimates on this slide):

[A single (typical) dataset generated from the three-leaf model]

Histograms (PRF vs RF)



Histograms of individual tree estimates $\mathbb{E}[\widehat{Y_{new}|x_{new}}]$'s for $x_{new}=(-0.3, 0.2)$. The vertical dashed black line indicates the true value $\mu_{new} = \mathbb{E}[Y_{new}|x_{new}]=-0.2$.

[Summary over R=300 repeated datasets]

Table: MSEs (PRF vs RF)

n	PRF	RF	ratio
1000	0.0270	0.0878	0.3075
2000	0.0117	0.0536	0.2183
3000	0.0056	0.0368	0.1522
4000	0.0024	0.0292	0.0822
5000	0.0025	0.0286	0.0874
6000	0.0019	0.0255	0.0745
7000	0.0010	0.0262	0.0382
8000	0.0009	0.0253	0.0356

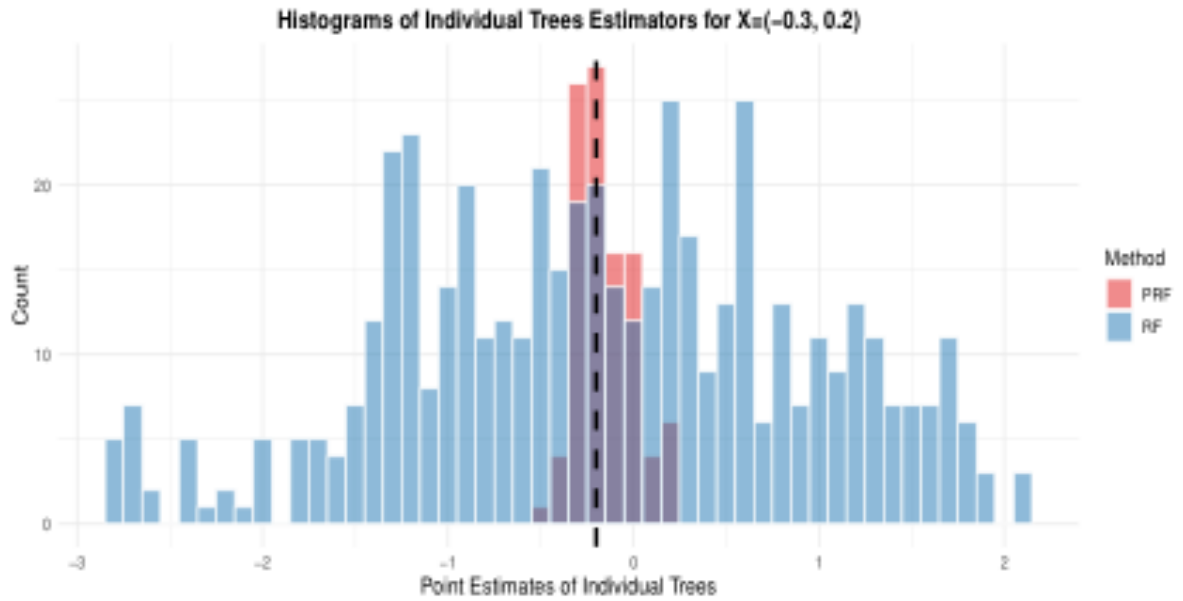
MSE $= \frac{1}{R}\sum_{r=1}^{R}\{\mathbb{E}(Y_{new}|x_{new}) - \widehat{Y_r}\}^2$ where in rth repeat $\widehat{Y_r}$ is average of individual tree estimates $\mathbb{E}[\widehat{Y_{new}|x_{new}}]$'s for $x_{new}=(-0.3, 0.2)$

Example Three: Principled Ensemble Forest

Why is it a principled ensemble forest? Because it is much more focused

- Artificial randomness not by sub-sampling/bootstrapping, but by mimicking u^{rel}
- Smaller number of trees required; concentrated around truth

Histograms (PRF vs RF)



Histograms of individual tree estimates $\widehat{\mathbb{E}[Y_{new}|x_{new}]}$'s for $x_{new}=(-0.3, 0.2)$.
The dashed black line indicates the true value $\mathbb{E}[Y_{new}|x_{new}]=-0.2$.

[fitting a single (typical) dataset in the toy example setup]

Table: MSEs (PRF vs RF)

n	PRF	RF	ratio
1000	0.0270	0.0878	0.3075
2000	0.0117	0.0536	0.2183
3000	0.0056	0.0368	0.1522
4000	0.0024	0.0292	0.0822
5000	0.0025	0.0286	0.0874
6000	0.0019	0.0255	0.0745
7000	0.0010	0.0262	0.0382
8000	0.0009	0.0253	0.0356

$\text{MSE}=\frac{1}{R}\sum_{r=1}^{R}\{\mathbb{E}(Y_{new}|x_{new})-\widehat{Y_r}\}^2$ where in rth repeat
$\widehat{Y_r}$ is ave of individual tree estimates $\widehat{\mathbb{E}[Y_{new}|x_{new}]}$'s
for $x_{new}=(-0.3, 0.2)$ [Repeat:R=300 datasets]

We discussed in this talk –

- Statistics and data science began about 250 years ago with the concept of "**inverse probability**."
- Classical approaches rely on **probabilistic models** and **analytical derivations** of the likelihood function (and its variants).
- Modern inferential approaches leverage **generative models** and **artificial datasets** to develop statistical inference and uncertainty quantification.

■ **For biopharmaceutical and clinical research, the central challenge remains the same in the ML/AI era:** How can we continue to develop reliable methods for quantifying uncertainty in treatment effects, biomarkers, and AI-assisted decision-making?

Open discussion: Questions are welcome!

Thank You!



References

- Dalmasso, Nicole et al. (2024). *Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning for Reliable Simulator-Based Inference*. arXiv: 2107.03920 [stat.ME]. URL: https://arxiv.org/abs/2107.03920.
- Hou, Xiaotian et al. (2025). "Repro Samples Method for Model-Free Inference in High-Dimensional Binary Classification". In: *arXiv preprint arXiv:2510.01468*.
- Tomaselli, Luca, Valerio Ventura, and Larry Wasserman (2025). *Robust Simulation-Based Inference*. arXiv: 2508.02404 [stat.ME]. URL: https://arxiv.org/abs/2508.02404.
- Wang, Peng, Min-ge Xie, and Linjun Zhang (2026). "Finite- and Large Sample Inference for Model and Coefficients in High-Dimensional Linear Regression with Repro Samples". In: *The Annals of Statistics* 54.2, pp. 834–860. DOI: 10.1214/25-AOS2591.
- Xie, Minge and Peng Wang (2026). *Repro Samples Method for a Performance Guaranteed Inference in General and Irregular Inference Problems*. arXiv: 2402.15004 [stat.ME]. URL: https://arxiv.org/abs/2402.15004.