A Bayesian Decision Theory Paradigm for Test & Evaluation

Jim Ferry

DATAWorks 2023
Alexandria, VA
April 27, 2023

Purpose and Overview

- **Purpose**: Develop a Bayesian version of T&E
  - Leverage expert inputs to represent prior knowledge about system being tested
  - Update knowledge with test data
  - Formulate utility functions to represent requirements and other stakeholder priorities
  - Provide test recommendations that optimize expected utility of testing vs. cost of testing
- **Overview**:
  - Classical statistics vs. Bayesian reasoning
  - Bayesian Decision Theory (BDT)
  - BDT paradigm for T&E
  - Decision charts
  - Excursion: raw distances vs. hit/miss data
  - Summary

Classical Statistics vs. Bayesian Reasoning

- Classical statistics… Bayesian reasoning… what's the difference?

- With infinite data, not much!
  - Let's flip a coin forever
  - HHTHHTTTTTHHHHHTTTHTTHHHTTHHHHTHHTTHHHHHHTTHTTTTHHHHTTTHTTHH...
  - The fraction of H's goes to a limiting value of $0.618034\ldots$
  - Statistics and Bayesian reasoning agree: the coin's probability of being heads is $p = 0.618034\ldots$

- What about with finite data?
  - Example 1: first 60 trials have 33 H and 27 T
  - Example 2: first 6 trials have 4 H and 2 T

Classical Statistics vs. Bayesian Reasoning

- Classical Statistics
  - Estimate p from data: $\hat{p} = \frac{h}{n}$
  - Compute quality of solution:
    - p in this range consistent with observing h:

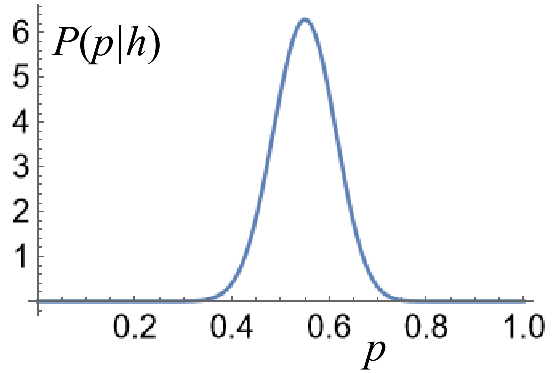
$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$ (Wald interval)

  - Example 1: $n = 60, h = 33$
    - $\hat{p} = 0.55$
    - Confidence Interval $= [0.4241, 0.6759]$
  - Example 2: $n = 6, h = 4$
    - $\hat{p} = 0.6667$
    - Confidence Interval $= [0.2895, 1.0439]$

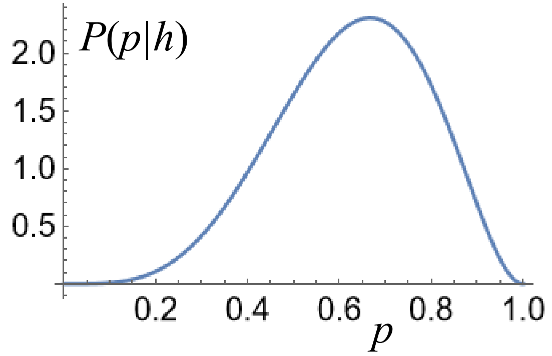
(Wald doesn't apply for small n)

- Bayesian Reasoning
  - Elicit prior belief $P(p)$ about p
    - E.g., $P(p) = 1$ (uniform on $[0,1]$)
  - Update prior belief to posterior $P(p|h)$.

Example 1: $n = 60, h = 33$



Example 2: $n = 6, h = 4$



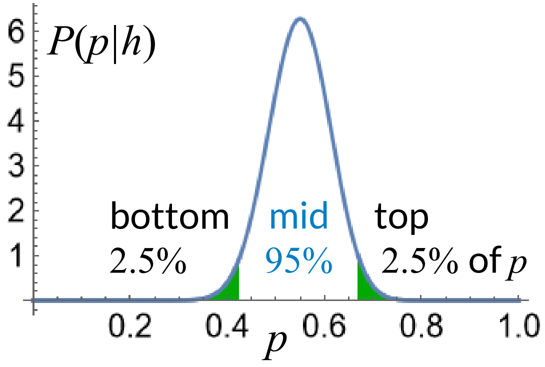
Classical Statistics vs. Bayesian Reasoning

- Classical Statistics
  - p unknown, but not *random* – no model for p
  - 95% *confidence intervals*
    - For any p in confidence interval,
      a guarantee that the observed h
      falls in the middle 95% of outcomes
  - Example 1: $n = 60, h = 33$ ✓ [32 | mid 95% of h | 46]
    $p = 0.65$
    - $\hat{p} = 0.55$
    - Confidence Interval $= [0.4241, 0.6759]$
  - Example 2: $n = 6, h = 4$
    - $\hat{p} = 0.6667$
    - Confidence Interval $= [0.2895, 1.0439]$

- Bayesian Reasoning
  - p is a random variable, so a prior necessary
    - A distribution on p is <u>always available</u>
    - More data = less dependence on prior
  - 95% *containment intervals* for $P(p|h)$

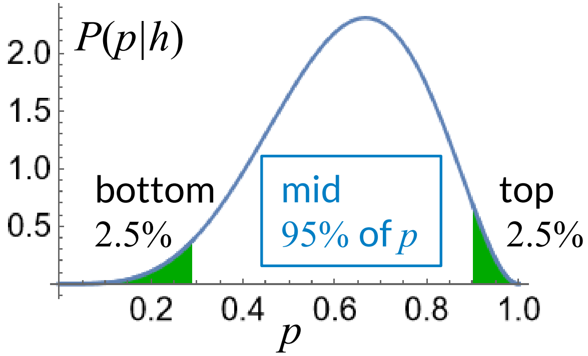
$[0.4245, 0.6694]$

Example 1: $n = 60, h = 33$



$[0.2904, 0.9010]$

Example 2: $n = 6, h = 4$



Classical Statistics vs. Bayesian Reasoning

- Classical Statistics
  - Pearson (1894) and Fisher (1925) [1,2]
  - Algorithmic mindset
    - Process data into an estimate of truth
    - Determine how much confidence one should have in the estimates
  - Provides a large set of tools for processing data and interpreting results
    - Easier to apply than Bayesian reasoning
    - But hard to interpret for complex problems

[1] K. Pearson, "Contributions to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society A*, **185**, 71-110, 1894.
[2] R.A. Fisher, *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd, 1925.
[3] R.T. Cox, *The Algebra of Probable Inference*, Johns Hopkins University Press, 1961.
[4] E.T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, 2003.

- Bayesian Reasoning
  - Increasingly widespread since 1990s
    - MCMC and VI computational methods
  - Scientific mindset
    - Focus on causal mechanisms by which truth causes the data to occur
  - Bayesian reasoning: the *unique* extension of classical logic to handle uncertainty [3,4]
  - Harder to apply: requires
    - Distilling key factors that drive behavior of data rather than selecting tools to apply
    - Representing and maintaining probability distributions, rather than computing numbers

Bayesian Decision Theory (BDT)

- Bayesian reasoning: why put in the effort?
  - Modeling causal mechanisms incorporates expert scientific knowledge
  - Maintaining probability distributions is the logically correct way to manage uncertainty
  - A probability distribution <u>always available</u> enables a *killer app*: Bayesian Decision Theory [5]
- Bayesian Decision Theory (BDT)
  - Distills stakeholder priorities into a *utility function* that defines how good a system is
  - Utility function quantifies cost/benefit of Accepting a system given
    - Some quantification of the uncertainty about its performance characteristics
    - The operational environment in which the system is required to perform
  - Utility function + probability distribution over system behavior:
    - Can make optimal decisions about how to test system… accounting for *cost* of tests

[5] L.C. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer, 2013.

BDT Framework for T&E: Spiral 1

- **Outcomes**
  - x
- Context
- Parametrization
- Utility
- Decisions

- In Spiral 1 T&E framework, a system is used repeatedly
  - Each use produces an *outcome* x
  - Testing requires outcomes to be known
- First step of framework: map test results to outcomes x
  - Test results can contain unstructured material: text, etc.
  - Map unstructured material into structured form for analysis
- Examples of outcomes:
  - x = hit/miss
  - x = hit/miss + if miss, failure stage that caused miss
  - x = error in meters
  - x = detect/non-detect + if detected, error in meters

BDT Framework for T&E: Spiral 1

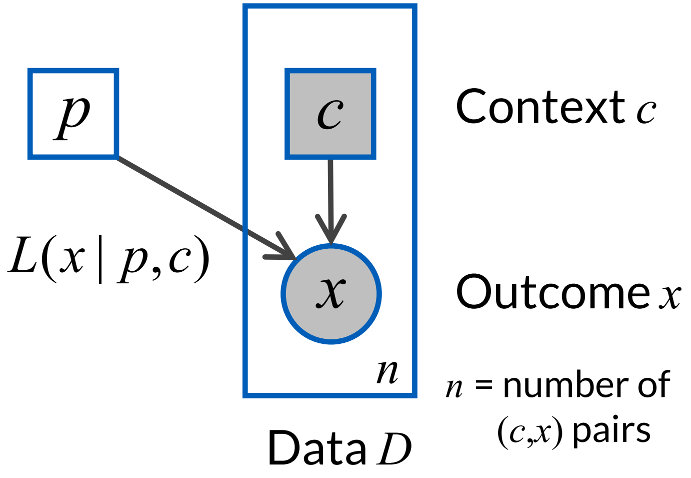
- Outcomes
  - x
- **Context**
  - c
- Parametrization
- Utility
- Decisions

- Outcomes influenced by (known) *context* c
- Context can include
  - Categorical data
    - Type of round, type of target, etc.
  - Specified environmental conditions
    - Range, depth, angle to horizon, angle to target, angle to sun
  - Unintentional-but-measured environmental conditions
    - Wind speed, temperature, etc.
    - Note: can be used in modeling outcomes x, but not in planning test design
  - Timestamp of test
    - To use for temporal correlations of unmeasured variables
  - Configuration of system
    - Various internal parameter settings

BDT Framework for T&E: Spiral 1

- Outcomes
  - x
- Context
  - c
- **Parametrization**
  - p
- Utility
- Decisions

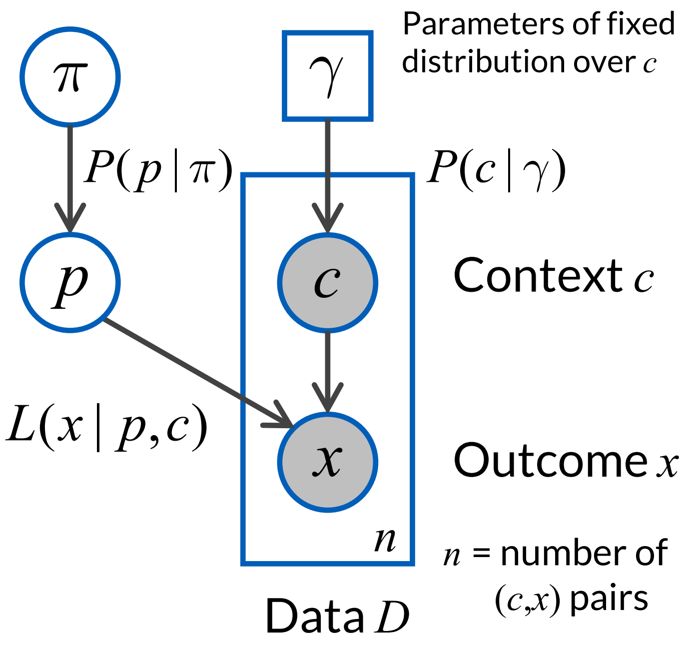
- System represented by (unknown) *parameter vector* p
- Model of outcomes: $L(x|p,c)$
  - $L(x|p,c)$ = probability of x given parameter vector p and context c
- Thought experiment: $L(x|p,c)$ as simulator
  - For given p, run simulator on each c
  - For each c, produce histogram of x
- With infinite data, could find true p
  - Would be an excellent model of system
  - Very useful for operational planning
- With finite data… estimate p?
- No: update prior over p to posterior over p



BDT Framework for T&E: Spiral 1

- Outcomes
  - x
- Context
  - c
- Parametrization
  - p
- **Utility**
  - U
- Decisions

- Define family $P(p|\pi)$ of probability distributions over p
  - Begin with prior $P(p|\pi_0)$, update to posterior $P(p|\pi_D)$ based on data D
- Define *utility* $U(\pi)$ of Accepting system given knowledge π
  - Example: "compliance utility"
    - $U_C(\pi) \doteq \begin{cases} U_1 & \text{if } \pi \text{ "good"} \\ -C_0 & \text{otherwise} \end{cases}$



- In general, utility may depend on
  - True parameter vector p
  - Context c in which system used
- But roll up into a metric depending only on knowledge π
  - E.g., $U(\pi) = \mathbb{E}_c\left[\mathbb{E}_p\left[U(\pi, p; c)\right]\right]$

BDT Framework for T&E: Spiral 1

- Outcomes
  - x
- Context
  - c
- Parametrization
  - p
- Utility
  - U
- **Decisions**
  - d

- Test event: max of n tests, then final decision required
- At *decision points*, pick T&E *actions* that yield optimal results
  - Example of actions: {A,R,T} = Accept/Reject system or continue to Test
  - Example of decision points: make decision $d \in$ {A,R,T} after every test
- Utilities: $U(\pi)$ for Accept, 0 for Reject, and each Test costs c_T
- $u(x_{1:k})$ = utility of outcomes being $x_{1:k}$ after k tests
- Backward recursion generates optimal decisions d:

$$\boxed{u(x_{1:k}) = \max\left\{U(\pi_k), 0, \mathbb{E}_{x_{k+1}}\left[u(x_{1:k+1})\right] - c_T\right\}}$$

Utility of d = Accept

Utility of d = Reject

Utility of d = Test → **Can compute because we know**

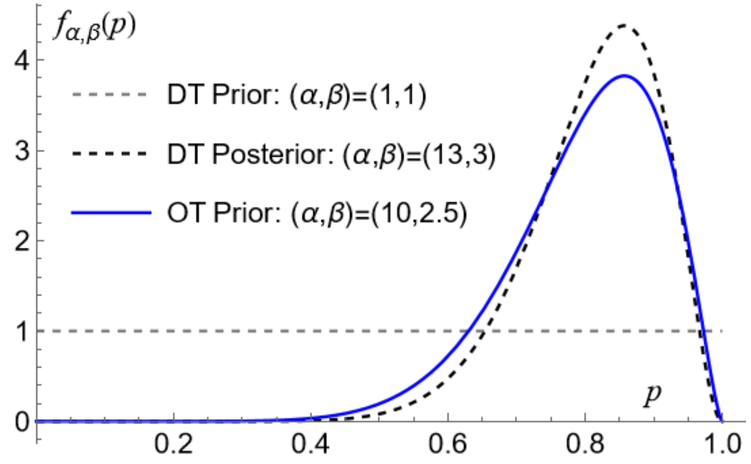
$$\begin{cases} L(p \mid \pi_k) \\ L(x_{k+1} \mid p, c_{k+1}) \end{cases}$$

Distribution of parameter vector given data thus far

Distribution of next outcome given parameter vector

Example: Hit/Miss Data

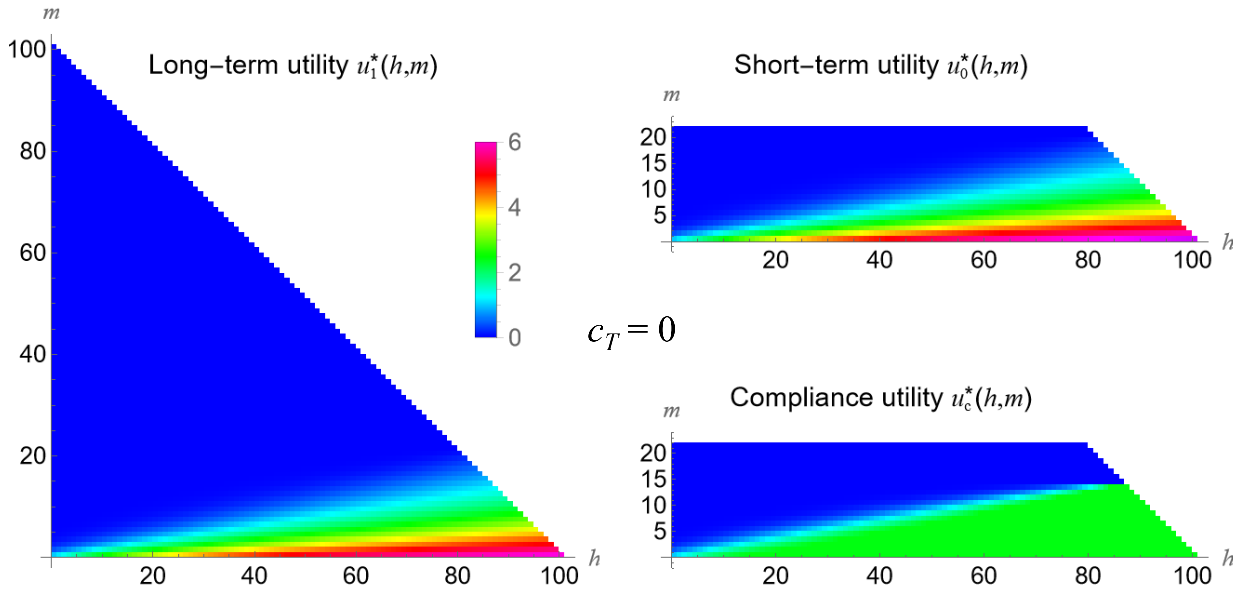
- Conduct up to n = 100 hit/miss trials ("coin flips")
  - Unknown value of p = P(hit)
  - $\pi = (\alpha,\beta)$: $P(p|\pi) = \mathrm{B}(p;\alpha,\beta)$ (beta distribution over p)
  - Prior $P(p|\pi_0) = \mathrm{B}(p;10,2.5)$



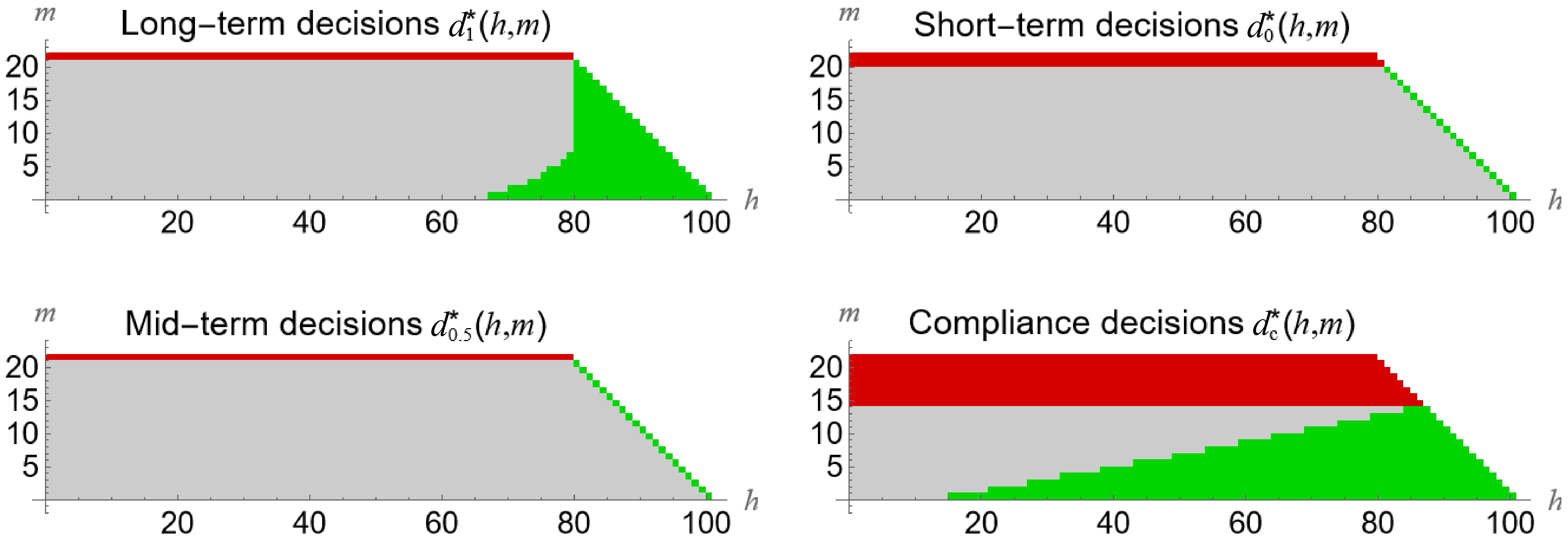
- Various Acceptance utilities $U(\pi)$
  - Long-term: based on true p only
  - Short-term: based on knowledge π
  - Mid-term: long/short compromise
  - Compliance:

$$U_C(\pi) \propto \begin{cases} 1 & \text{if } P(p \geq 0.8) \geq 95\% \\ -0.3 & \text{otherwise} \end{cases}$$

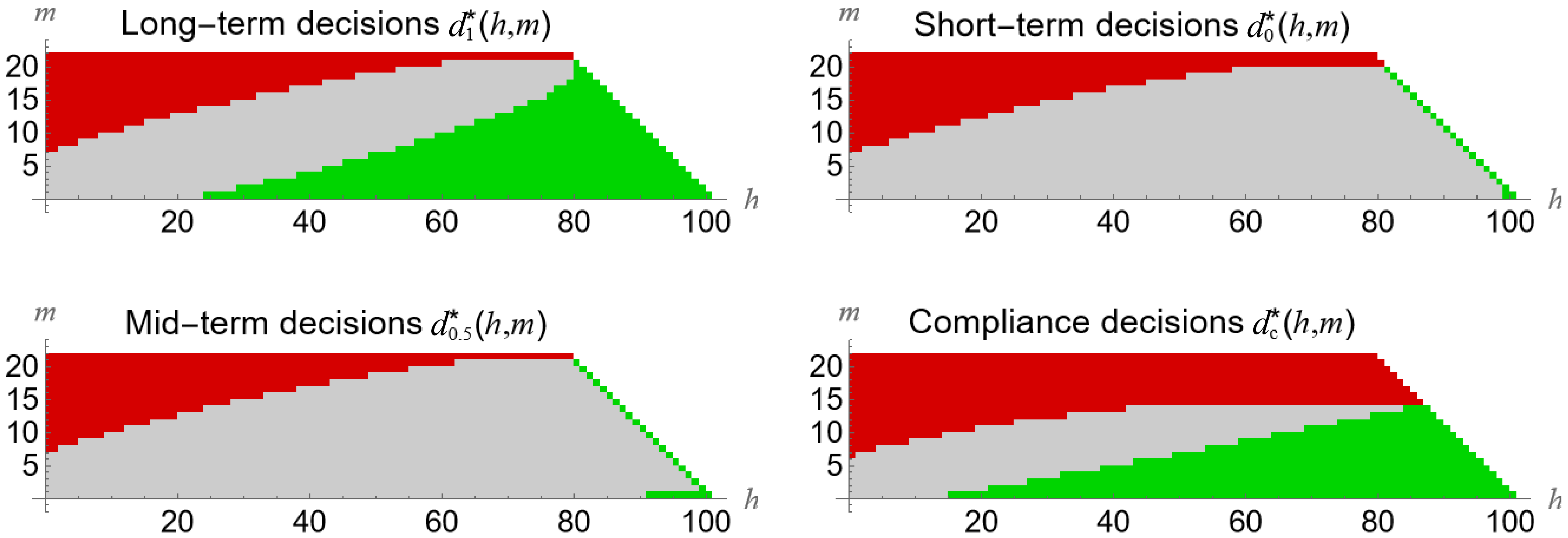
- Focus on *decision charts*



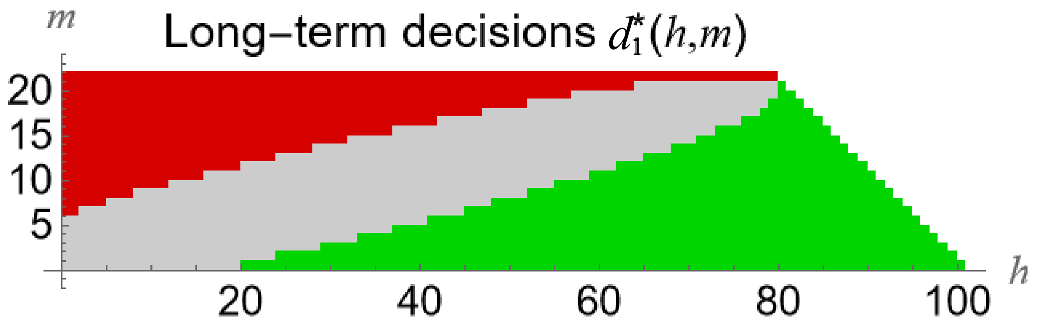
Decision Charts for $c_T = 0$

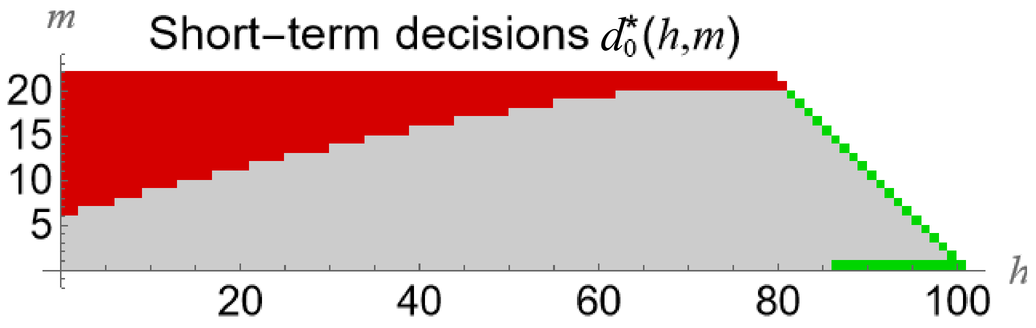


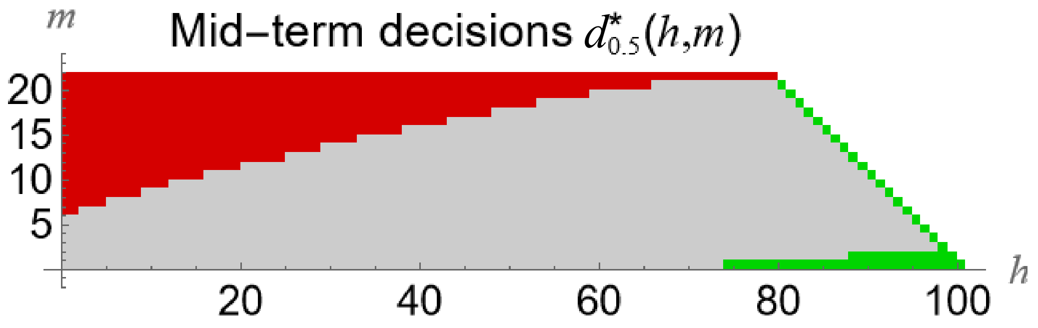
Decision Charts for c_T = 0.0001

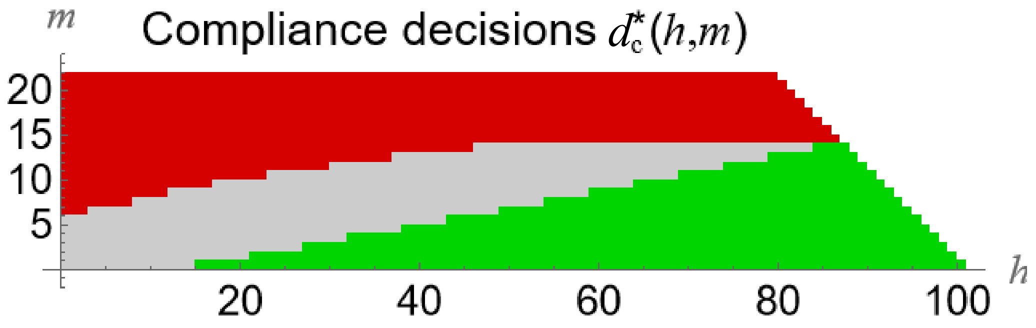


Decision Charts for c_T = 0.0003

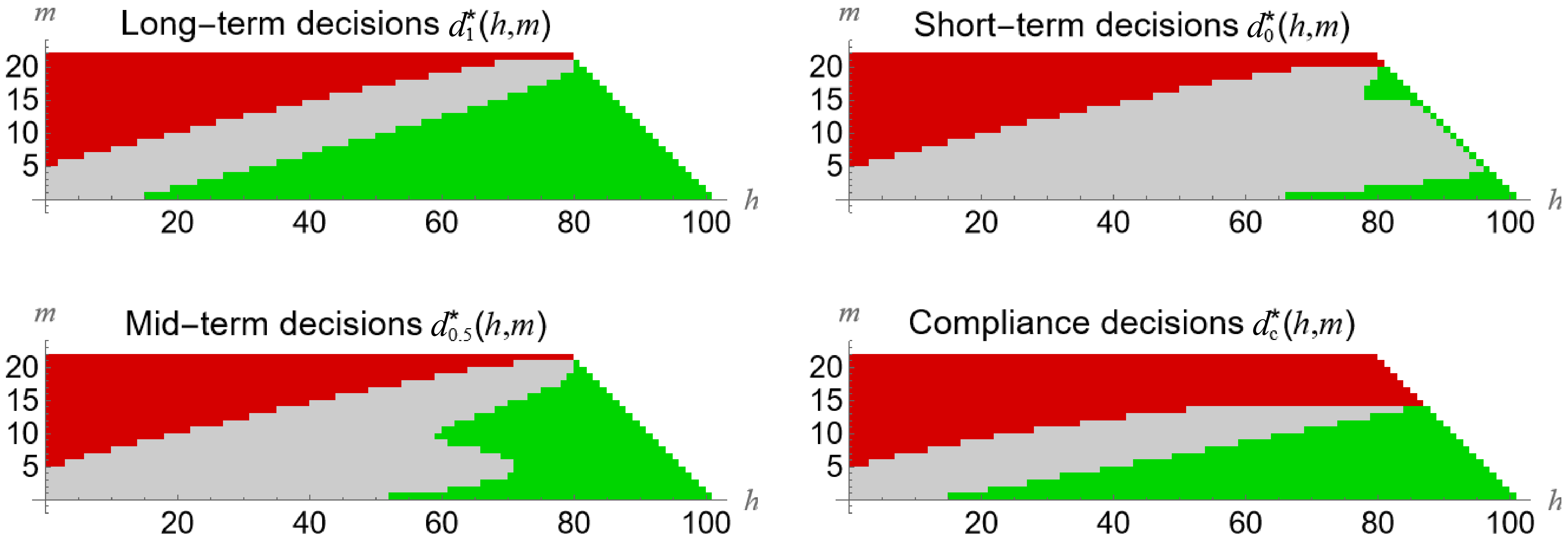




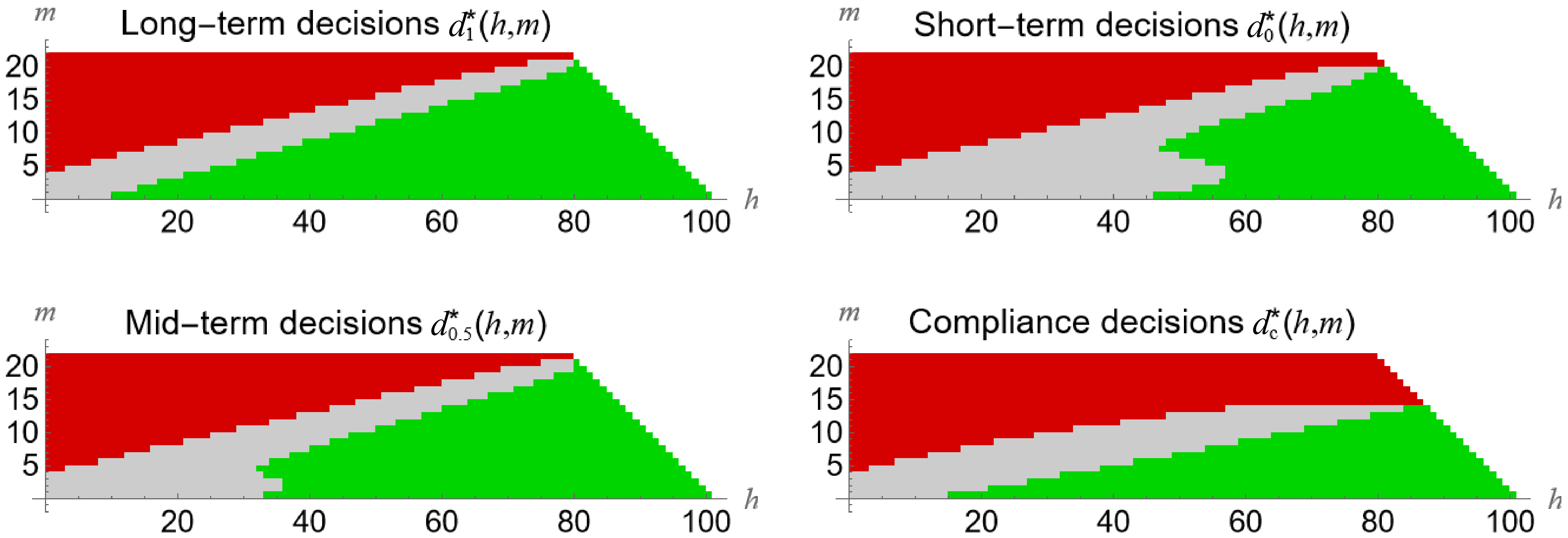




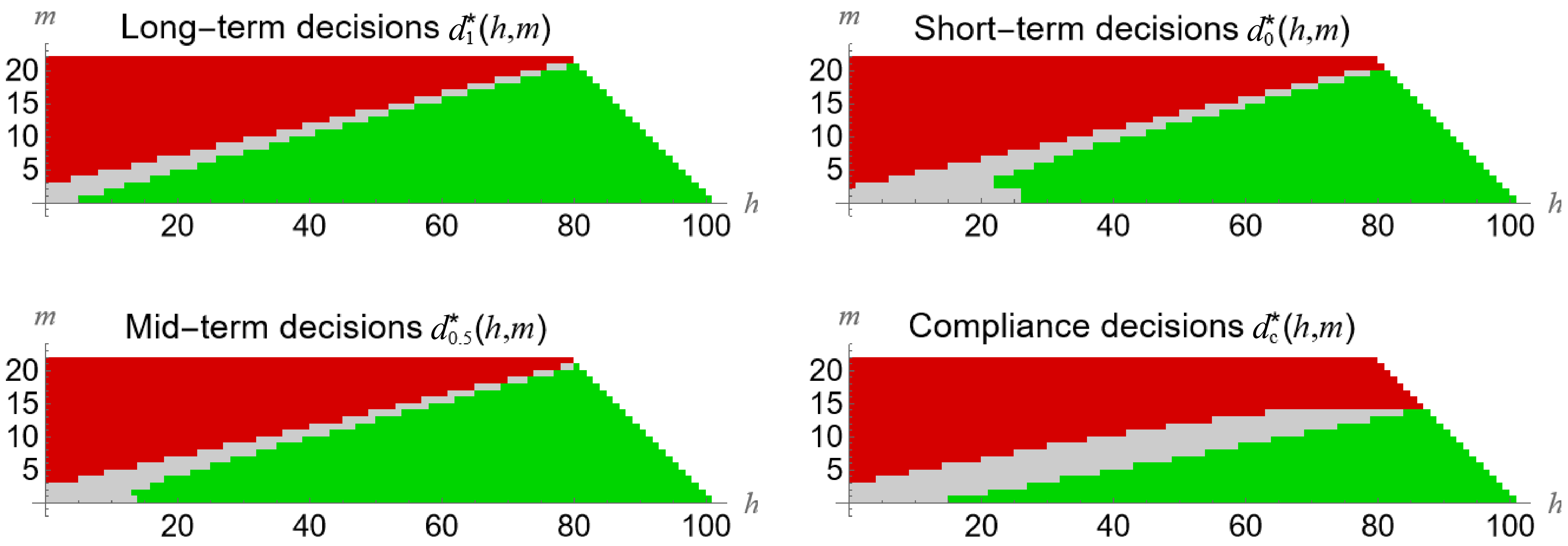
Decision Charts for $c_T = 0.001$



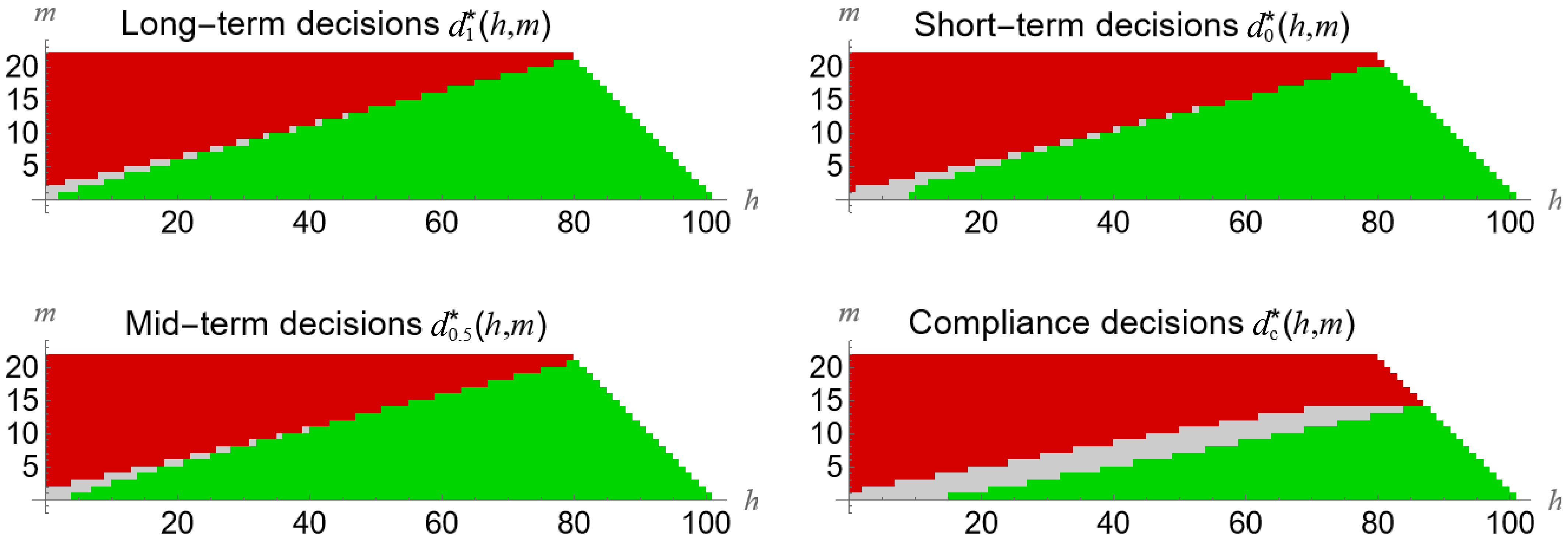
Decision Charts for $c_T = 0.003$



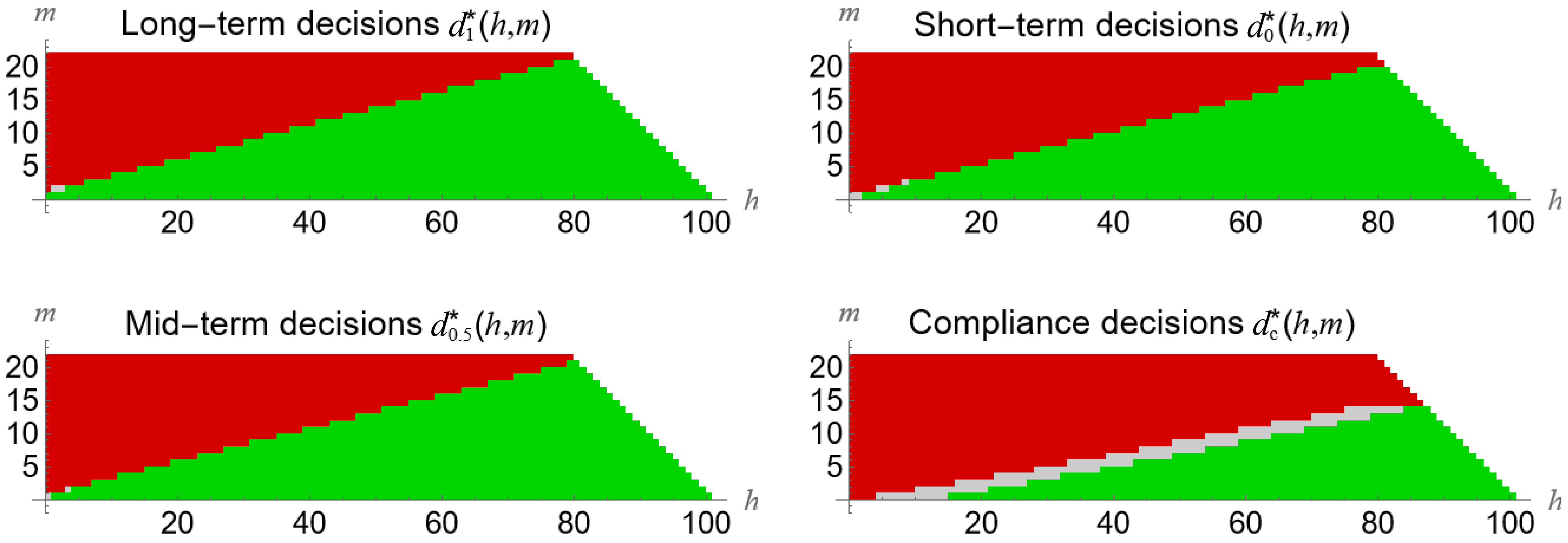
Decision Charts for $c_T = 0.01$



Decision Charts for $c_T = 0.03$

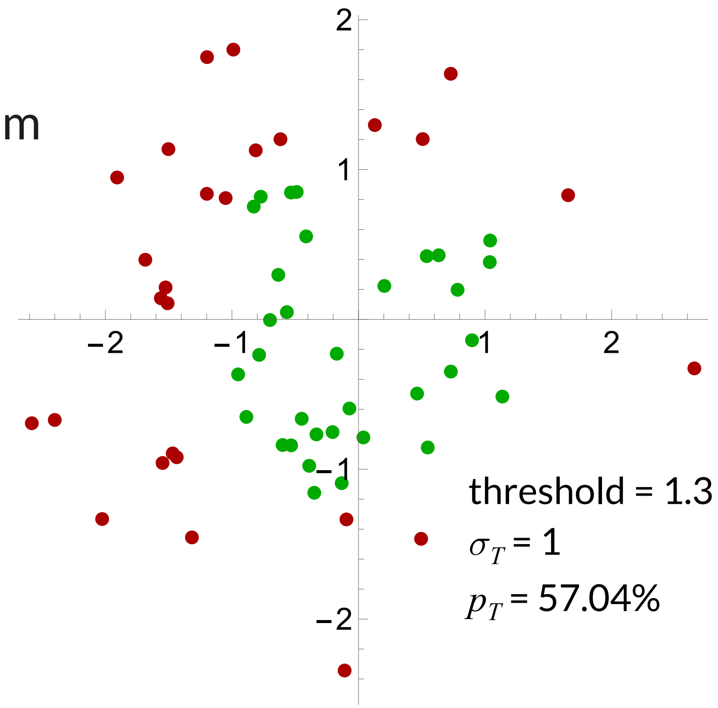


Decision Charts for $c_T = 0.1$



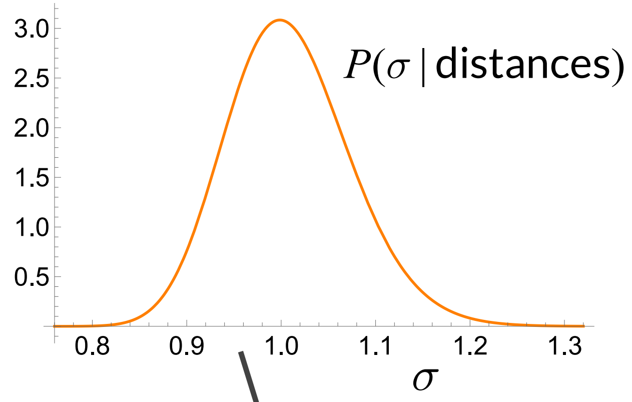
Hit/Miss vs. Continuum Measurements

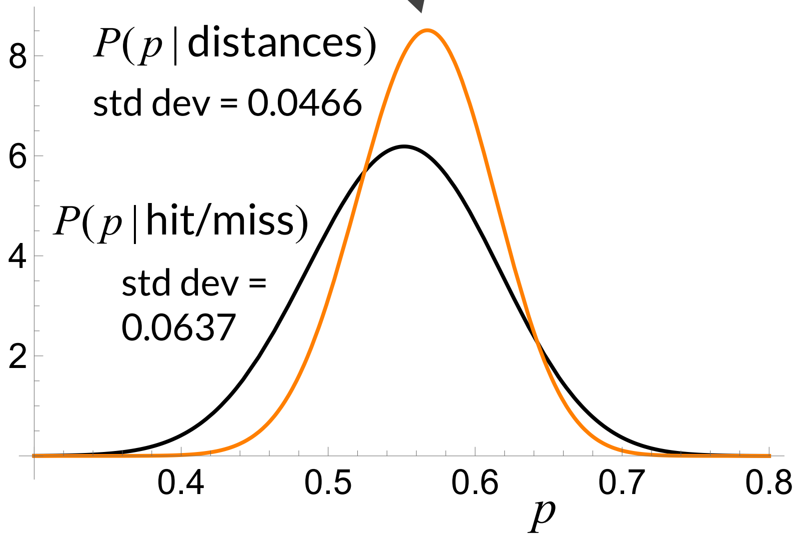
- In practice, a hit/miss call is often derived from a *distance*
  - E.g., a hit is called when the distance to a target is below some given threshold
- Isn't better to use the raw distances to estimate system performance?
  - Yes: one throws away information in the conversion to hit/miss
  - However: this requires understanding the distance distribution
    - In particular, *outliers* can disrupt inference procedures that ignore them
- How much does using raw distances help?
  - Idealized model: some ground-truth 2-d Gaussian distribution
    - Squared distances are exponentially distributed in this case
  - Consider a fixed hit/miss distance threshold
    - True distribution has some (unknown) hit probability p_T
  - What does inference do in the hit/miss and continuum cases?



Hit/Miss vs. Continuum Measurements

- Compute posterior distribution on σ using raw distances
- Convert to distribution on p
- Compare to posterior on p using hit/miss data
  - Standard deviation = 0.0466 using distance data
  - Standard deviation = 0.0637 using hit/miss data
- Does this hold in general?
  - Find formulas for average variance of p in each case
    - As a function of n and p_T
  - Given n and p_T, consider average variance using distances
    - How many times larger does n have to be to get same average variance using hit/miss data?
  - Find asymptotic result as $n \to \infty$





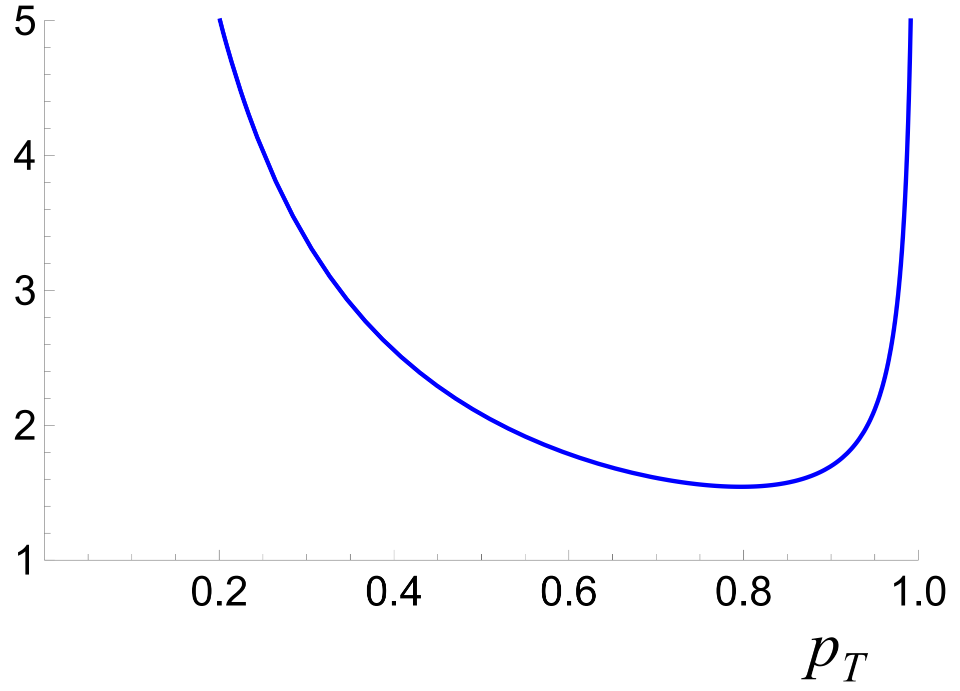
Hit/Miss vs. Continuum Measurements: Result

- Nice exact formula!
  - For 2-d Gaussian case

$$\frac{p_T}{(1-p_T)\log^2(1-p_T)}$$

- Each distance observation worth at least 1.544 hit/miss observations
  - Minimum occurs at $p_T = 0.797$

Number of hit/miss observations each distance observation is worth



Summary

- Test & Evaluation (T&E) is important, but increasingly complex
  - Essential to developing effective, reliable systems
  - How does one do this in a cost-effective manner?
- Bayesian reasoning
  - Models what systems are (p) in terms of probability distributions
    - Over the outcomes x they deliver
    - In the range of operational contexts c required
  - Can update probability distribution over p given data
- Bayesian Decision Theory
  - Captures stakeholder priorities in *utility function*
  - Utility function + probability distribution over system behavior = optimal decision-making for T&E, including cost of testing