Statistical Decision Theory

Perry Williams

Department of Fish, Wildlife, and Conservation Biology
Department of Statistics
Colorado State University

26 June 2016

Motivation, History, and Fundamentals

Why Statistical Decision Theory?

- Decisions are made at every step in scientific investigation
  - Data collection
  - Model selection
  - Summary statistics
  - Management

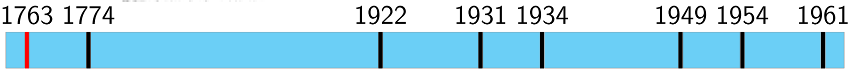
Why Statistical Decision Theory?

- Decisions are made at every step in scientific investigation
  - Data collection
  - Model selection
  - Summary statistics
  - Management
- SDT provides a cohesive framework for decision making
  - Data collection–Dynamic adaptive sampling
  - Model selection–Optimal prediction
  - Summary statistics–Bayes rules
  - Management actions–Optimal management

Rev. T. Bayes

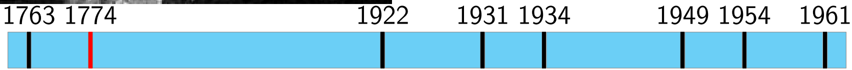
Bayes' Theorem appeared in *"An Essay Towards Solving a Problem in the Doctrine of Chances"*

"Aldrich suggests that we interpret [Bayes' definition of probability] in terms of expected utility, and thus that Bayes' result would make sense only to the extent to which one can bet on its observable consequences."
-Stephen Fienberg, 2006.



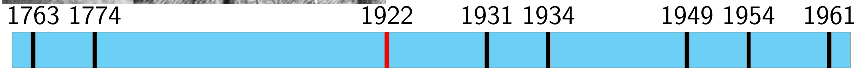
Laplace published "Memoire sur la Probabilité des Causes par las Évènements"

- Elaborate example of inverse probability
- Uniform prior distributions
- Methods for choosing estimators that minimize posterior loss



Fisher published "On the Mathematical Foundations of Theoretical Statistics"

- Rejected inverse probability
- Grounded his theory on frequency interpretation of probability
- Obviated the need for prior distributions
- Introduced likelihood
- Tests of significance

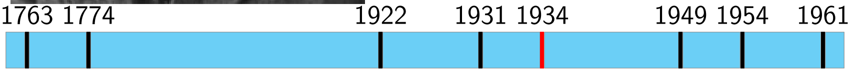


Fisher on Probability and Decisions

> "We aim, in fact, at methods of inference which should be equally convincing to all rational minds, irrespective of any intentions they may have in utilizing the knowledge inferred."

1763 1774 1922 1931 1934 1949 1954 1961

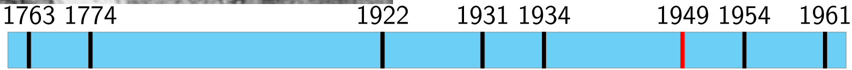
History



History

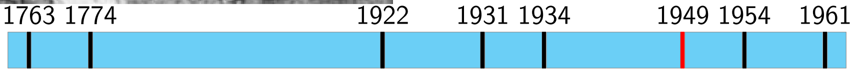
Wald published "Statistical Decision Functions"

- Unified statistical theory by treating statistical problems as special cases of zero-sum two-person games
- Statistical inference was viewed as a special case of decision theory (*c.f.*, Von Neumann and Morgenstern 1944)



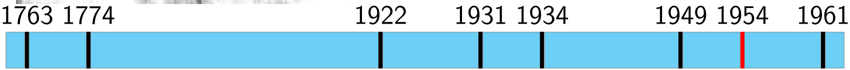
"It is well recognized that the statistical estimation theory should and can be organized within the framework of the theory of statistical decision functions (Wald 1950)"

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle.

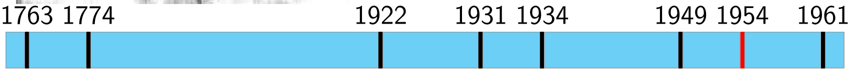


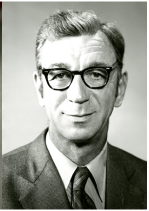
Savage published "The Foundations of Statistics"

- Set the stage for Bayesian revival



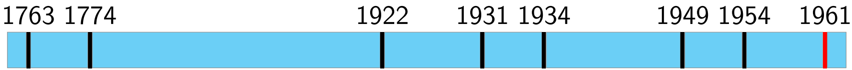
"Decision theory is the best and most stimulating, if not the only, systematic model of statistics."





Raiffa and Schlaifer published "Applied Statistical Decision Theory"

- Methods of Fisher, Neyman, and Pearson did not address the main problem of a businessman: how to make decisions under uncertainty
- Developed Bayesian decision theory



- F.P. Ramsey
- B. De Finetti
- J.M. Keynes
- H. Jeffreys
- D.V. Lindley
- D.R. Cox
- J.W. Tukey
- A. Birnbaum
- M. Kendall

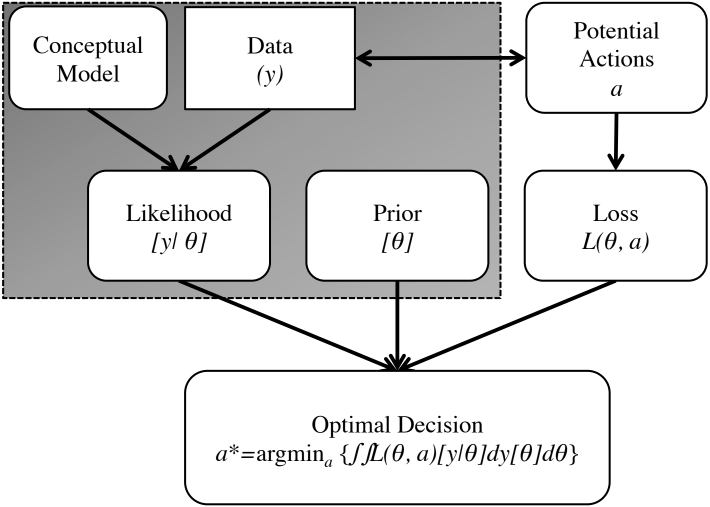
Fundamentals of SDT

Inferential Steps

1. Identify possible states of nature (support)
2. Assign prior probabilities
3. Assign model processes (for potentially many models)
4. Apply Bayes theorem to obtain posterior probabilities from data

Decision Steps

1. Identify possible states of nature (support)
2. Assign prior probabilities
3. Assign model processes
4. Apply Bayes theorem to obtain posterior probabilities from data
5. Enumerate possible decisions
6. Assign a loss function
7. Choose decision that minimizes expected loss



Williams, P.J., and M.B. Hooten. 2016. Combining statistical inference and decisions in ecology. Ecological Applications.

Basic Elements

Θ: potential states of nature

θ: true state of nature

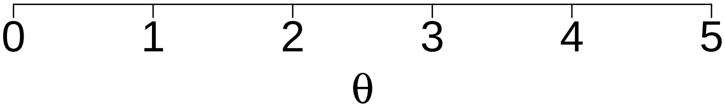
$\mathscr{A}$: potential actions

a: a specific action, possibly a function of data (i.e., $\delta(y)$)

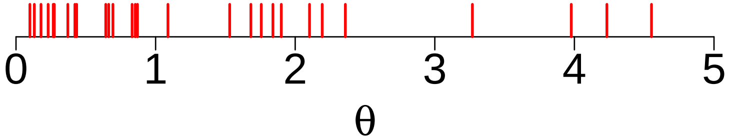
$L : \Theta \times \mathscr{A} \mapsto \mathbb{R}$: loss function

$L(\theta, a)$: loss incurred if action a is made and θ is true

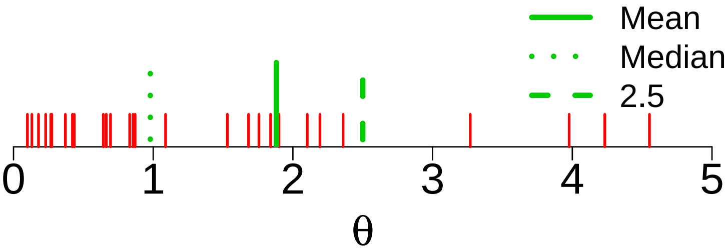
Example: Measurement With Uncertainty



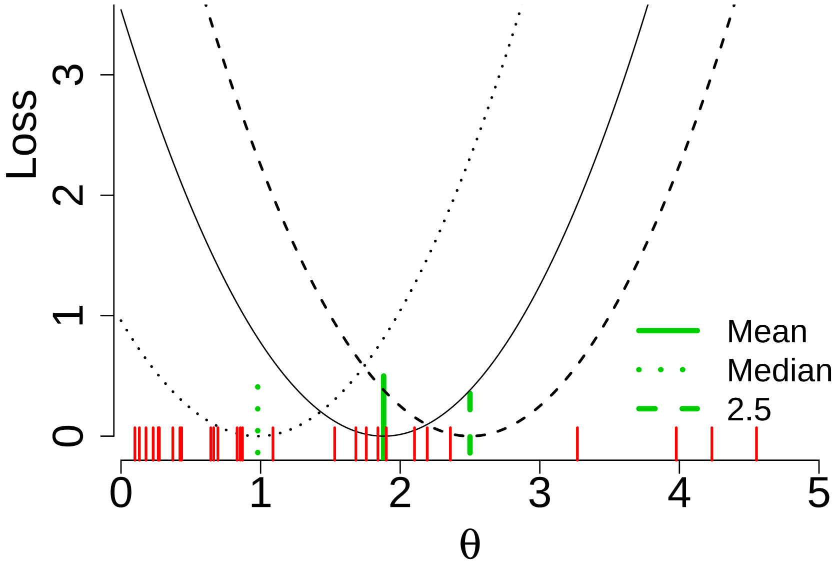
Example: Measurement With Uncertainty



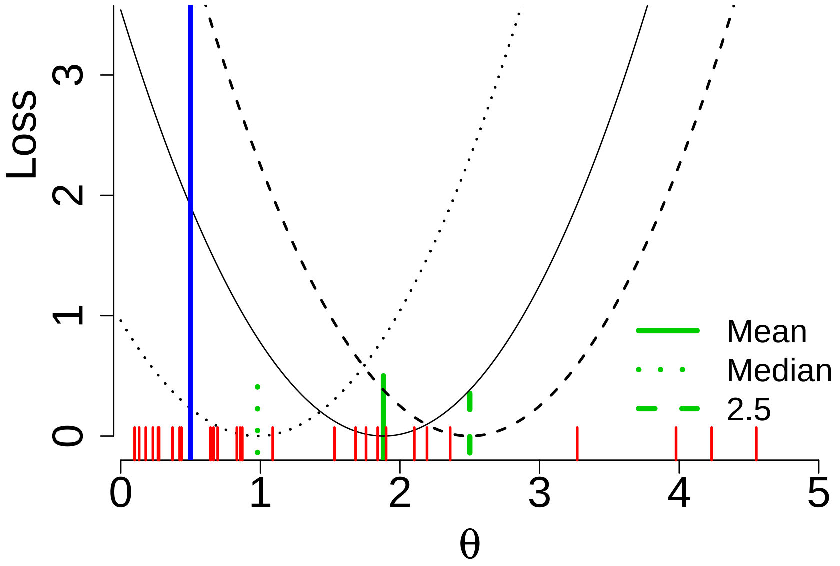
Example: Measurement With Uncertainty



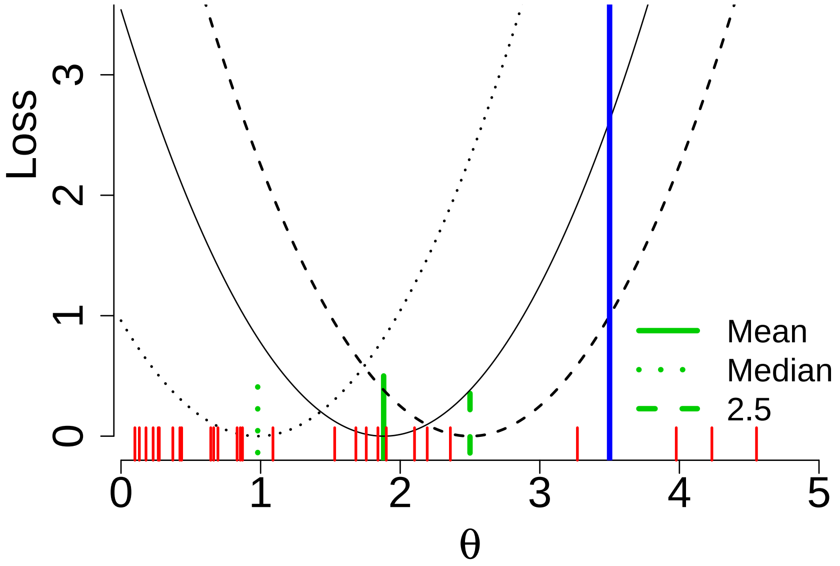
Example: Measurement With Uncertainty



Example: Measurement With Uncertainty



Example: Measurement With Uncertainty



Types of Loss

- Squared Error Loss:

$$L(\theta, a) = (\theta - a)^2$$

- Linear Loss:

$$L(\theta, a) = \begin{cases} c_1(\theta - a) \text{ if } \theta > a \\ c_2(a - \theta) \text{ if } \theta < a \end{cases}$$

- 0–1 Loss:

$$L(\theta, a) = \begin{cases} 0 \text{ if } \theta = a \\ 1 \text{ if } \theta \neq a \end{cases}$$

Risk

- Frequentist Risk
- Bayesian Expected Loss
- Bayesian Risk

Decision Rule $\delta(y)$

Y: a random variable that depends on θ

$\mathscr{Y}$: the sample space of Y

y: a realization from $\mathscr{Y}$

$\delta : \mathscr{Y} \mapsto \mathscr{A}$
(for any possible realization $y \in \mathscr{Y}$, δ describes which action to take)

Decision Rule Example

$$H_0 : \theta \geq 0$$
$$H_a : \theta < 0$$

$$\delta_1(y) = \begin{cases} a_1 = \text{Reject } H_0 & \bar{y} < -0.3 \\ a_2 = \text{Fail to reject } H_0 & \bar{y} \geq -0.3 \end{cases}$$
$$\delta_2(y) = \begin{cases} a_1 = \text{Reject } H_0 & \bar{y} < -0.5 \\ a_2 = \text{Fail to reject } H_0 & \bar{y} \geq -0.5 \end{cases}$$

Frequentist Risk

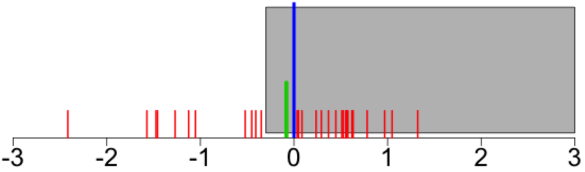
- How much you expect to lose when using a decision rule $\forall y \in \mathscr{Y}$

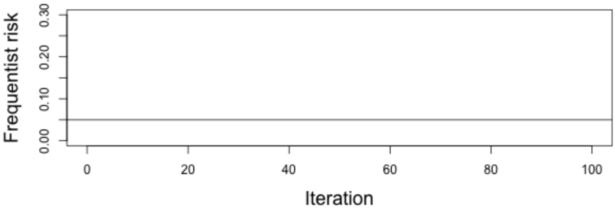
$$
\begin{aligned}
R(\theta, \delta) =& E[L(\theta, \delta(y))] \\
=& \int_{\mathscr{Y}} L(\theta, \delta(y))[y|\theta] dy
\end{aligned}
$$

Frequentist Risk and p-values

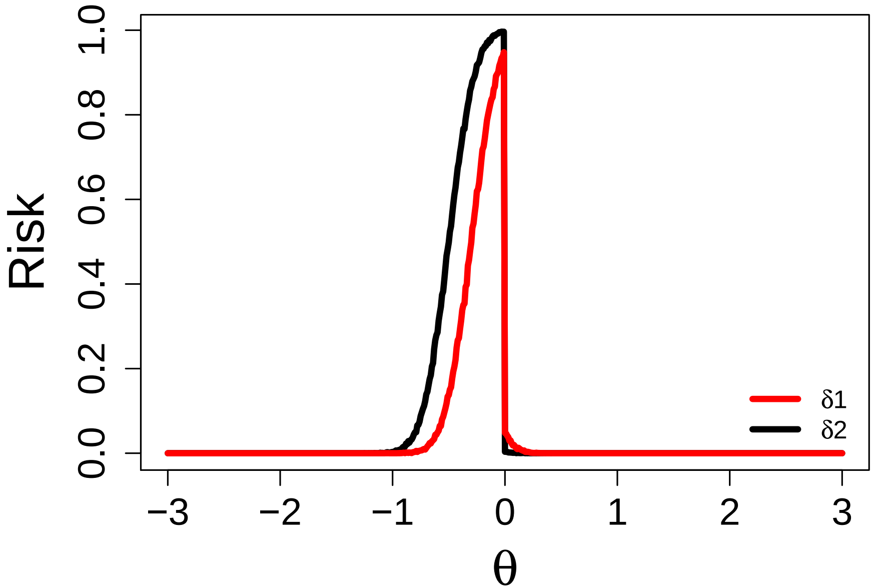
$$\delta_1(y) = \begin{cases} a_1 = \text{Reject } H_0 & \bar{y} < -0.3 \\ a_2 = \text{Fail to reject } H_0 & \bar{y} \geq -0.3 \end{cases}$$

$$L(\theta, \delta(y)) = \begin{cases} 1 & \text{if reject } H_0 \text{ and } \theta \geq 0 \\ 1 & \text{if fail to reject } H_0 \text{ and } \theta < 0 \\ 0 & \text{otherwise} \end{cases}$$





Frequentist Risk



Decision Theory is Inherently Bayesian

- Goal of Decision Theory: Make a decision based on our belief in the probability of an unknown state
- Frequentist Probability: The limit of a state's relative frequency in a large number of trials
- Bayesian Probability: Degree of rational belief to which a state is entitled in light of the given evidence

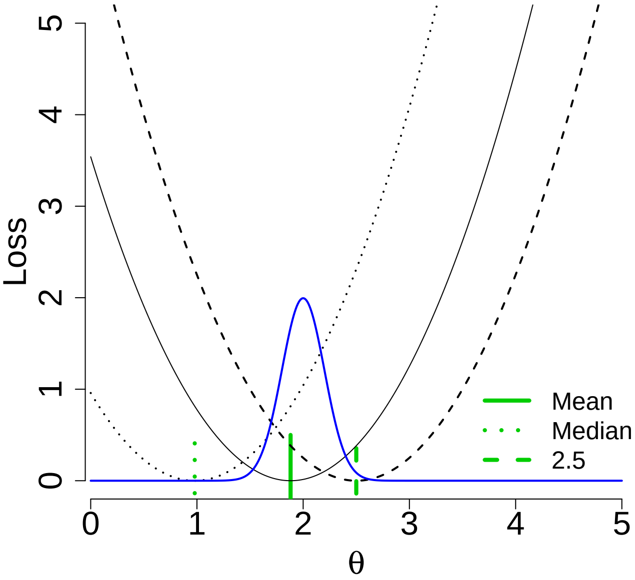
Bayesian Expected Loss

- Probability distribution assigned to θ
  - Prior probability distribution: $[\theta]$
  - Posterior probability distribution: $[\theta|y]$
- Bayesian Expected Loss: the loss averaged over the distribution of θ.

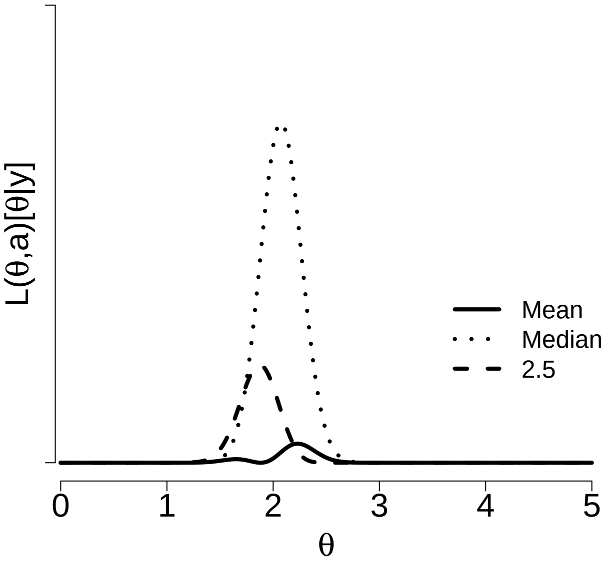
Bayesian Expected Loss

- $\rho(a) = \ E_{\theta} L(\theta, a) = \int_{\Theta} L(\theta, a)[\theta] d\theta$
- $\rho(a) = \ E_{\theta|y} L(\theta, a) = \int_{\Theta} L(\theta, a)[\theta|y] d\theta$
- Bayes Rule: $a^* = \text{argmin}_a(\rho(a))$

Bayesian Expected Loss



Bayesian Expected Loss



Bayes Risk

Mathematical Relationship Between Frequentist Risk and Bayes Expected Loss:

$$r(a) = \int_{\Theta} \left\{ \int_{\mathscr{Y}} L(\theta, a)[y|\theta]dy \right\}[\theta]d\theta$$

Note: $[y|\theta][\theta] = [\theta|y][y]$

$$r(a) = \int_{\mathscr{Y}} \left\{ \int_{\theta} L(\theta, a)[\theta|y]d\theta \right\}[y]dy$$

Bayesian Expected Loss vs. Frequentist Risk

- $R(\theta, \delta)$ integrates over y, but y is known (i.e., data)
- $R(\theta, \delta)$ is a function of unknown θ
- $R(\theta, \delta)$ doesn't make use of auxiliary information (e.g., prior knowledge)
- δ that minimizes $\rho(\delta)$ also minimizes $r(\delta)$ (don't need to integrate over hypothetical replicates of y)

Bayesian Point Estimation

Bayesian Point Estimation

- Suppose we want to summarize a posterior distribution $[\theta|y]$ with a point estimate
- We want to minimize the information lost using the point estimate to summarize $[\theta|y]$
- Is the choice of point estimate arbitrary?

Bayesian Point Estimation

- Suppose we want to summarize a posterior distribution $[\theta|y]$ with a point estimate
- We want to minimize the information lost using the point estimate to summarize $[\theta|y]$
- Is the choice of point estimate arbitrary?
- *Bayes Estimator*: Bayes Rule for point estimation

Bayes rule for squared-error loss

$$\rho(a) = \int_\theta (\theta - a)^2 [\theta|y] d\theta$$

$$\frac{d}{da}\rho(a) = 2E[\theta|y] - 2a$$

$$2E[\theta|y] - 2a \overset{\text{set}}{=} 0$$

$$a^* = E[\theta|y]$$

Implied Loss of Posterior Mean

$$a = E[\theta|y]$$

$$f''(a)(a - E[\theta|y]) = 0$$

$$\int f''(a)(a - E[\theta|y])da = \int (f'(a)(a - \theta) - f(a) + g(\theta))[\theta|y]d\theta$$

$$L(\theta, a) = f'(a)(a - \theta) - f(a) + g(\theta)$$

Bayes rule for absolute-error loss

$$\rho(a) = \int_{\Theta} |\theta - a|[\theta|y]d\theta$$

$$\frac{d}{da}\rho(a) = -P(\theta \geq [a|y]) + P(\theta \leq [a|y])$$

$$-P(\theta \geq [a|y]) + P(\theta \leq [a|y]) \stackrel{\text{set}}{=} 0$$

$$a^* = \text{ median}([\theta|y])$$

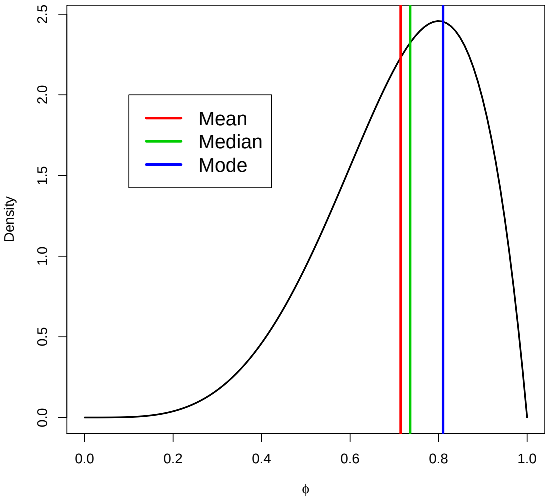
Bayes rule for 0–1 loss

$$\rho(a) = \int_{a_1}^{a_2} 0[\theta|y]d\theta + \int_{-\infty}^{a_1} 1[\theta|y]d\theta + \int_{a_2}^{\infty} 1[\theta|y]d\theta$$

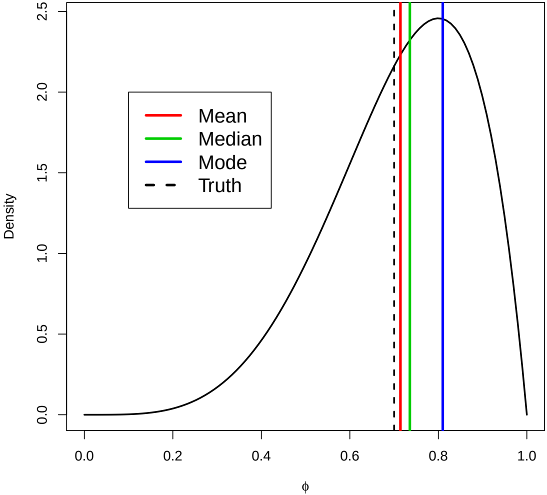
$$\frac{d}{da}\rho(a) = [a_1|y] - [a_2|y], \ \ as \ a_1 \rightarrow a \leftarrow a_2$$

$$a^* = \mathrm{mode}([\theta|y])$$

Asymmetric Distributions



Asymmetric Distributions

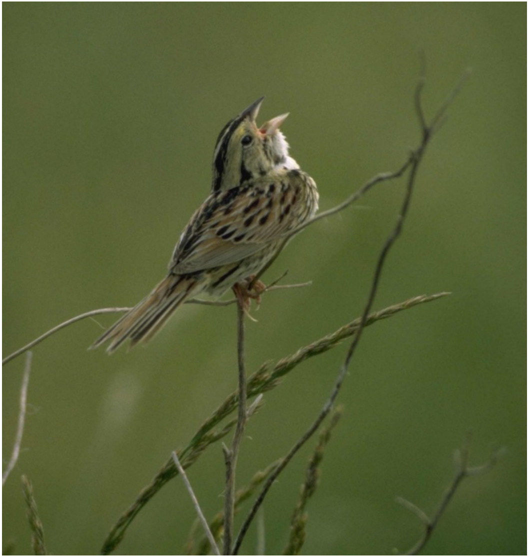


Bayesian Point Estimation

- Bayesian point estimation is *NOT* arbitrary!
- Different loss functions yield different point estimates.
  - SEL: Posterior mean
  - Absolute Loss: Posterior median
  - 0–1 Loss: Posterior mode
- A choice of point estimator implies decision maker's choice of loss function (or class of functions).

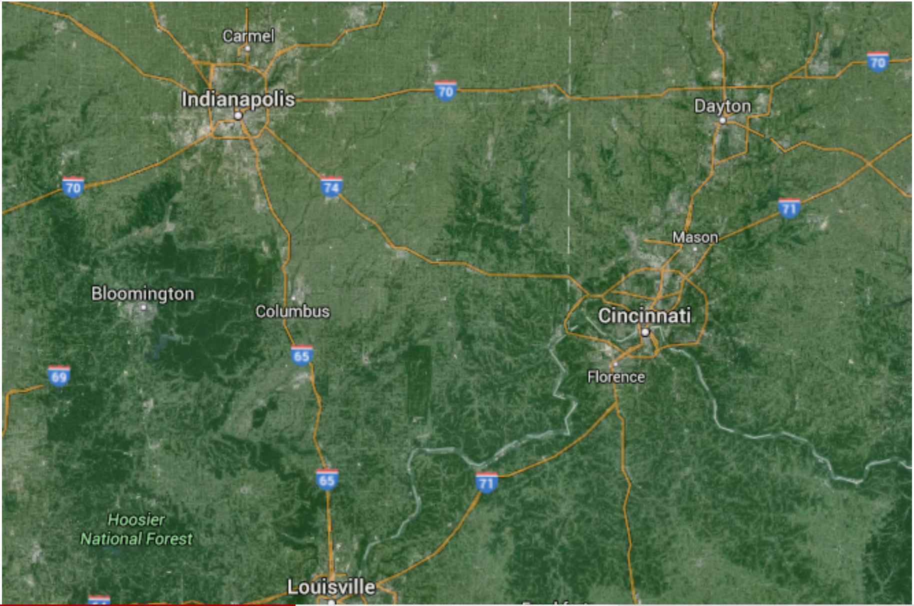
Management Decisions

Henslow's Sparrow



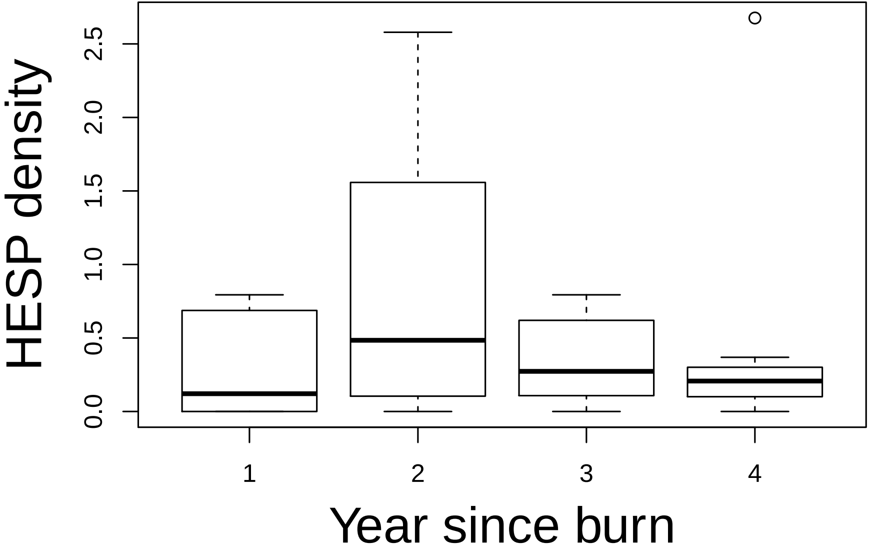
Williams, P.J., and M.B. Hooten. 2016. Combining statistical inference and decisions in ecology. Ecological Applications.

Big Oaks National Wildlife Refuge



Optimal Prescribed Fire Return Interval?

Non-Optimal Solution



Elements for SDT

- Data: Density estimates
- Prior Information: Mean henslow's sparrow densities at other sites (Herkert and Glass 1999)
- Loss Function related to cost and management objectives

Response-to-fire model

$$y_{j,t} \sim \text{Poisson}(A_j \lambda_{j,t}),$$

$$log(\lambda_{j,t}) = \boldsymbol{x'}_{j,t}\boldsymbol{\beta} + \eta_j,$$

$$\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\mu}, \sigma^2 I),$$

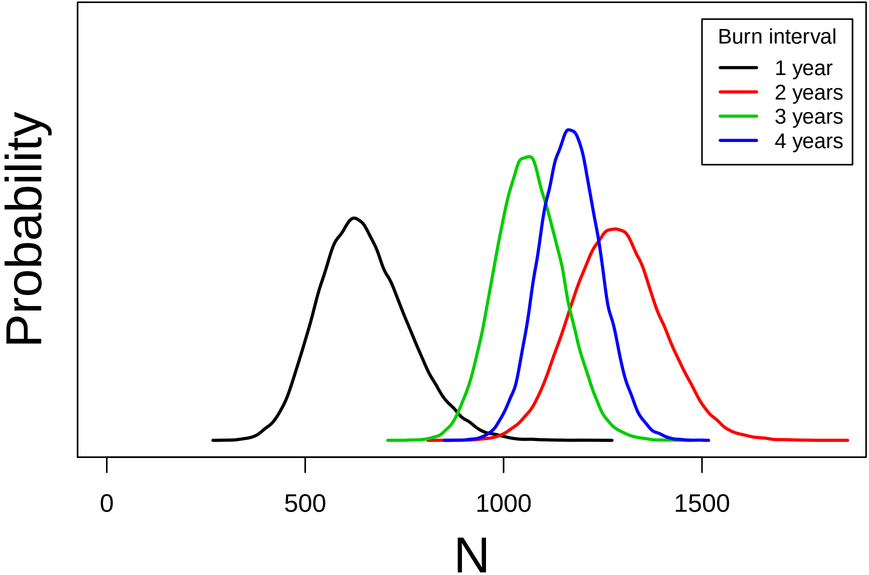
$$\eta_j \sim \text{Normal}(0, \sigma^2_\eta).$$

Cumulative abundance as a derived parameter

$$N_a = \lim_{\tilde{T} \to \infty} \frac{20 \sum_{j=1}^{8} \sum_{t=T+1}^{\tilde{T}} A_i \lambda_{j,t}}{\tilde{T} - T}$$

j=1, a=3	$A_1\lambda_{1,1}$	$A_1\lambda_{1,2}$	$A_1\lambda_{1,3}$	$A_1\lambda_{1,1}$	$A_1\lambda_{1,2}$	$A_1\lambda_{1,3}$	…	$A_1\lambda_{1,1}$	$A_1\lambda_{1,2}$	$A_1\lambda_{1,3}$
	t=1	*t=2*	*t=3*	*t=4*	*t=5*	*t=6*	…			*t=* $\tilde{T}$

Posterior Distributions

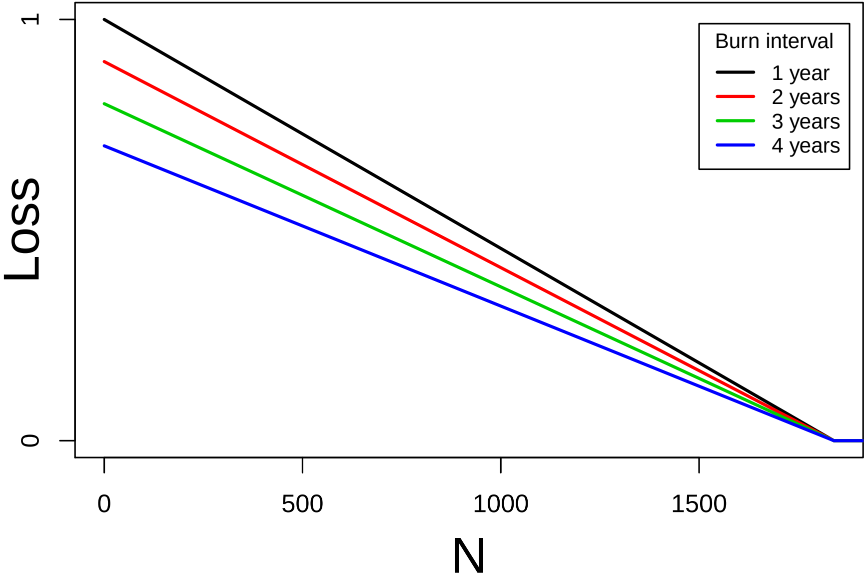


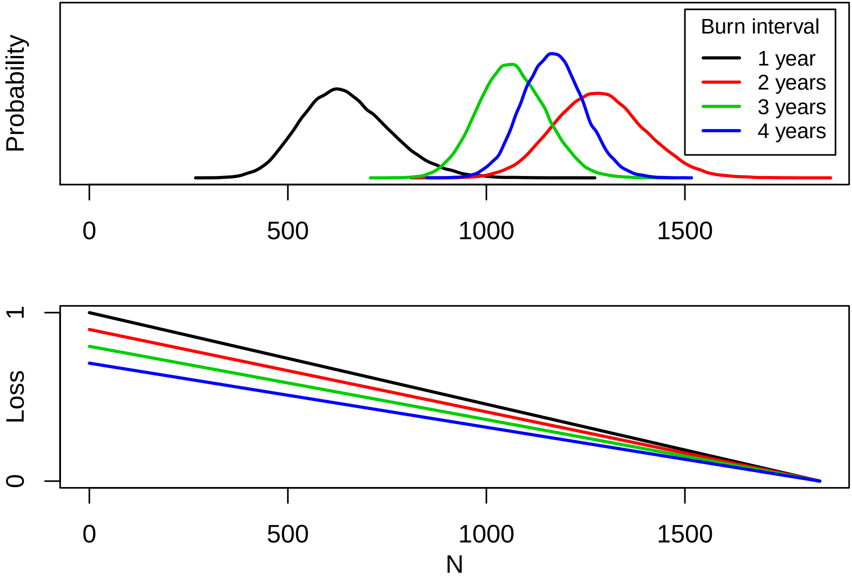
Axioms of Henslow's Sparrow Management

1. Frequent fire intervals are more expensive
2. More birds are better
3. The relative importance of cost decreases as abundance increases

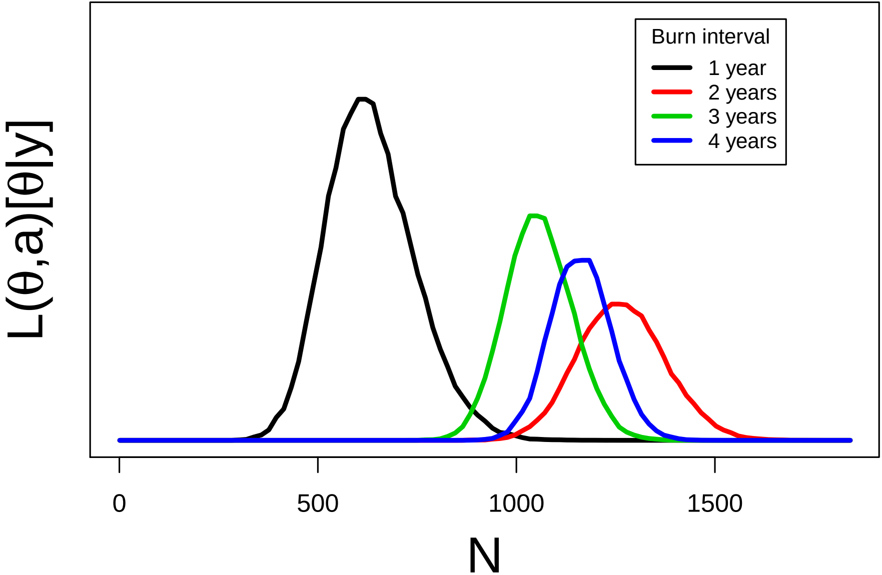
Henslow's sparrow loss function

$$L(\theta, a) = \begin{cases} \alpha_0(a) + \alpha_1(a) N_{a,\boldsymbol{\theta}}, & N_{a,\boldsymbol{\theta}} < 1835 \\ 0, & N_{a,\boldsymbol{\theta}} \geq 1835 \end{cases}$$

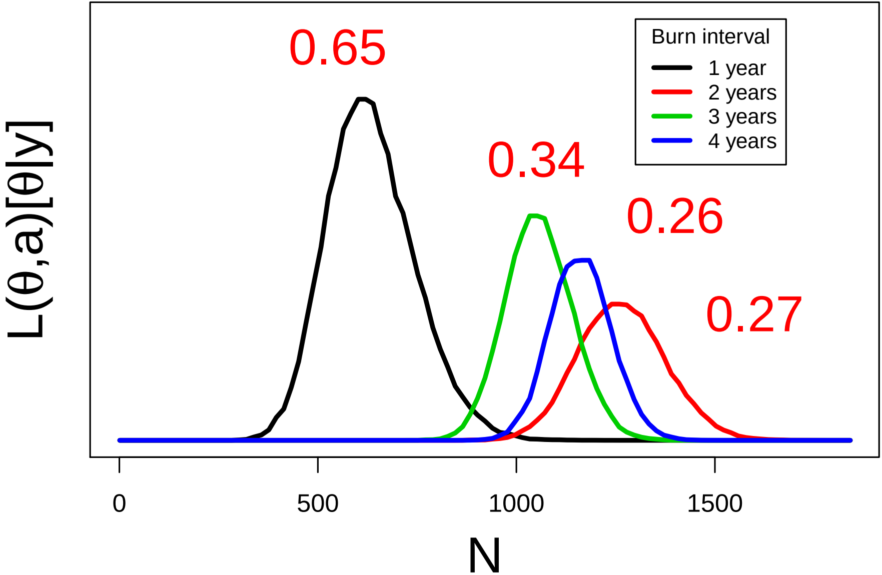

Burn interval
1 year
2 years
3 years
4 years
Loss
1
0
0
500
1000
1500
N


Probability
Burn interval
1 year
2 years
3 years
4 years
0
500
1000
1500
Loss
1
0
0
500
1000
1500
N

Convolution of Loss Function and Posterior Distribution



Bayes Expected Loss



Model Selection Example

- Posterior predictive distribution

$$[\tilde{\mathbf{y}}|\mathbf{y}, m] = \int [\tilde{\mathbf{y}}|\mathbf{y}, \boldsymbol{\theta}^{(m)}][\boldsymbol{\theta}^{(m)}|\mathbf{y}]d\boldsymbol{\theta}^{(m)}$$

- Posterior predictive loss (Gelfand and Ghosh 1998)

$$L(\tilde{\mathbf{y}}, \mathbf{y}, a) = L(\tilde{\mathbf{y}}, a) + kL(\mathbf{y}, a)$$

- Posterior predictive expectation of loss for model m, and action a.

$$E_{\tilde{\mathbf{y}}|\mathbf{y},m} L(\tilde{\mathbf{y}}, \mathbf{y}, a)$$

Summary

- SDT combines statistical analyses and decision theory
- Implications for data collection, point estimation, and model selection
- Ecological/Management applications