# The likelihood approach to statistical decision problems

Marco Cattaneo

Department of Mathematics
University of Hull

21 May 2015

# introduction

- the classical and Bayesian approaches to statistics are unified and generalized by the corresponding decision theories

# introduction

- the classical and Bayesian approaches to statistics are unified and generalized by the corresponding decision theories

- the likelihood approach to statistics is extremely successful in practice, but it is not unified and generalized by a decision theory

# introduction

- the classical and Bayesian approaches to statistics are unified and generalized by the corresponding decision theories
- the likelihood approach to statistics is extremely successful in practice, but it is not unified and generalized by a decision theory
- is such a likelihood decision theory possible?

# introduction

- the classical and Bayesian approaches to statistics are unified and generalized by the corresponding decision theories
- the likelihood approach to statistics is extremely successful in practice, but it is not unified and generalized by a decision theory
- is such a likelihood decision theory possible?
- in statistics, $L$ usually denotes:

# introduction

- the classical and Bayesian approaches to statistics are unified and generalized by the corresponding decision theories
- the likelihood approach to statistics is extremely successful in practice, but it is not unified and generalized by a decision theory
- is such a likelihood decision theory possible?

- in statistics, $L$ usually denotes:
  - likelihood function

# introduction

- the classical and Bayesian approaches to statistics are unified and generalized by the corresponding decision theories
- the likelihood approach to statistics is extremely successful in practice, but it is not unified and generalized by a decision theory
- is such a likelihood decision theory possible?

- in statistics, $L$ usually denotes:
  - likelihood function
  - loss function

# introduction

- the classical and Bayesian approaches to statistics are unified and generalized by the corresponding decision theories

- the likelihood approach to statistics is extremely successful in practice, but it is not unified and generalized by a decision theory

- is such a likelihood decision theory possible?

- in statistics, $L$ usually denotes:
  - likelihood function (here $\lambda$)
  - loss function (here $W$)

# introduction

- the classical and Bayesian approaches to statistics are unified and generalized by the corresponding decision theories

- the likelihood approach to statistics is extremely successful in practice, but it is not unified and generalized by a decision theory

- is such a likelihood decision theory possible?

- in statistics, $L$ usually denotes:
  - likelihood function (here $\lambda$)
  - loss function (here $W$)

- **statistical model**: $(\Omega, \mathcal{F}, P_\theta)$ with $\theta \in \Theta$ (where $\Theta$ is a nonempty set) and random variables $X_i : \Omega \to \mathcal{X}_i$

# loss function

- a statistical **decision problem** is described by a loss function

$$W : \Theta \times \mathcal{D} \to [0, +\infty),$$

  where $\mathcal{D}$ is a nonempty set

# loss function

- a statistical **decision problem** is described by a loss function

$$W : \Theta \times \mathcal{D} \to [0, +\infty),$$

  where $\mathcal{D}$ is a nonempty set

- intended as unification (and generalization) of statistical inference,

# loss function

- a statistical **decision problem** is described by a loss function

$$W : \Theta \times \mathcal{D} \to [0, +\infty),$$

  where $\mathcal{D}$ is a nonempty set

- intended as unification (and generalization) of statistical inference, in particular of:
  - point estimation (e.g., with $\mathcal{D} = \Theta$)
  - hypothesis testing (e.g., with $\mathcal{D} = \{H_0, H_1\}$)

# loss function

- a statistical **decision problem** is described by a loss function
$$W : \Theta \times \mathcal{D} \to [0, +\infty),$$
where $\mathcal{D}$ is a nonempty set

- intended as unification (and generalization) of statistical inference, in particular of:
    - point estimation (e.g., with $\mathcal{D} = \Theta$)
    - hypothesis testing (e.g., with $\mathcal{D} = \{H_0, H_1\}$)

- most successful general methods:
    - point estimation: maximum likelihood estimators
    - hypothesis testing: likelihood ratio tests

# loss function

- a statistical **decision problem** is described by a loss function

$$W : \Theta \times \mathcal{D} \to [0, +\infty),$$

  where $\mathcal{D}$ is a nonempty set

- intended as unification (and generalization) of statistical inference, in particular of:
  - point estimation (e.g., with $\mathcal{D} = \Theta$)
  - hypothesis testing (e.g., with $\mathcal{D} = \{H_0, H_1\}$)

- most successful general methods:
  - point estimation: maximum likelihood estimators
  - hypothesis testing: likelihood ratio tests

- these methods do not fit well in the setting of classical or Bayesian decision theory: here they are unified (and generalized) in likelihood decision theory

# simple decision example

- random sample of 3 black balls from an urn containing 100 balls, of which $\theta \in \Theta = \{50, 99, 100\}$ are black: select $d \in \mathcal{D} = \{$ "50", "not 50" $\}$

| $W$ | "50" | "not 50" | $\lambda$ |
|---|---|---|---|
| 50 | 0 | 15 | 0.12 |
| 99 | 1 | 0 | 0.97 |
| 100 | 1 | 0 | 1.00 |

# simple decision example

- random sample of 3 black balls from an urn containing 100 balls, of which $\theta \in \Theta = \{50, 99, 100\}$ are black: select $d \in \mathcal{D} = \{\text{“50”}, \text{“not 50”}\}$

| $W$ | “50” | “not 50” | $\lambda$ |
|---|---|---|---|
| 50 | 0 | 15 | 0.12 |
| 99 | 1 | 0 | 0.97 |
| 100 | 1 | 0 | 1.00 |

- classical decision: choose what it means to **repeat** the experiment, select the decision rule minimizing the (pre-data) expected loss, and apply it to the particular data

# simple decision example

- random sample of 3 black balls from an urn containing 100 balls, of which $\theta \in \Theta = \{50, 99, 100\}$ are black: select $d \in \mathcal{D} = \{\text{“50”}, \text{“not 50”}\}$

| $W$ | “50” | “not 50” | $\lambda$ |
|---|---|---|---|
| 50 | 0 | 15 | 0.12 |
| 99 | 1 | 0 | 0.97 |
| 100 | 1 | 0 | 1.00 |

- classical decision: choose what it means to **repeat** the experiment, select the decision rule minimizing the (pre-data) expected loss, and apply it to the particular data (difficult and indirect)

# simple decision example

- random sample of 3 black balls from an urn containing 100 balls, of which $\theta \in \Theta = \{50, 99, 100\}$ are black: select $d \in \mathcal{D} = \{$ "50", "not 50" $\}$

| $W$ | "50" | "not 50" | $\lambda$ |
|---|---|---|---|
| 50 | 0 | 15 | 0.12 |
| 99 | 1 | 0 | 0.97 |
| 100 | 1 | 0 | 1.00 |

- classical decision: choose what it means to **repeat** the experiment, select the decision rule minimizing the (pre-data) expected loss, and apply it to the particular data (difficult and indirect)

- Bayesian decision: choose a **prior** on $\Theta$ and select the decision minimizing the (post-data) expected loss

# simple decision example

- random sample of 3 black balls from an urn containing 100 balls, of which $\theta \in \Theta = \{50, 99, 100\}$ are black: select $d \in \mathcal{D} = \{$ "50", "not 50" $\}$

| $W$ | "50" | "not 50" | $\lambda$ |
|---|---|---|---|
| 50 | 0 | 15 | 0.12 |
| 99 | 1 | 0 | 0.97 |
| 100 | 1 | 0 | 1.00 |

- classical decision: choose what it means to **repeat** the experiment, select the decision rule minimizing the (pre-data) expected loss, and apply it to the particular data (difficult and indirect)

- Bayesian decision: choose a **prior** on $\Theta$ and select the decision minimizing the (post-data) expected loss (prior dependent)

# simple decision example

- random sample of 3 black balls from an urn containing 100 balls, of which $\theta \in \Theta = \{50, 99, 100\}$ are black: select $d \in \mathcal{D} = \{\text{“50”}, \text{“not 50”}\}$

| $W$ | “50” | “not 50” | $\lambda$ |
|---|---|---|---|
| 50 | 0 | 15 | 0.12 |
| 99 | 1 | 0 | 0.97 |
| 100 | 1 | 0 | 1.00 |

- classical decision: choose what it means to **repeat** the experiment, select the decision rule minimizing the (pre-data) expected loss, and apply it to the particular data (difficult and indirect)

- Bayesian decision: choose a **prior** on $\Theta$ and select the decision minimizing the (post-data) expected loss (prior dependent)

- maximum likelihood decision: select the decision minimizing the loss when $\theta = \hat{\theta}_{ML}$ (i.e., “not 50”, since $\hat{\theta}_{ML} = 100$)

# simple decision example

- random sample of 3 black balls from an urn containing 100 balls, of which $\theta \in \Theta = \{50, 99, 100\}$ are black: select $d \in \mathcal{D} = \{$ "50", "not 50" $\}$

| $W$ | "50" | "not 50" | $\lambda$ |
|---|---|---|---|
| 50 | 0 | 15 | 0.12 |
| 99 | 1 | 0 | 0.97 |
| 100 | 1 | 0 | 1.00 |

- classical decision: choose what it means to **repeat** the experiment, select the decision rule minimizing the (pre-data) expected loss, and apply it to the particular data (difficult and indirect)
- Bayesian decision: choose a **prior** on $\Theta$ and select the decision minimizing the (post-data) expected loss (prior dependent)
- maximum likelihood decision: select the decision minimizing the loss when $\theta = \hat{\theta}_{ML}$ (i.e., "not 50", since $\hat{\theta}_{ML} = 100$) (too optimistic)

# likelihood function

- $\lambda_x : \Theta \to [0, 1]$ is the (relative) likelihood function given $X = x$, when

$$\sup_{\theta \in \Theta} \lambda_x(\theta) = 1 \quad \text{and} \quad \lambda_x(\theta) \propto P_\theta(X = x)$$

# likelihood function

- $\lambda_x : \Theta \to [0, 1]$ is the (relative) likelihood function given $X = x$, when

$$\sup_{\theta \in \Theta} \lambda_x(\theta) = 1 \quad \text{and} \quad \lambda_x(\theta) \propto P_\theta(X = x)$$

(with $\lambda_x(\theta) \propto f_\theta(x)$ as approximation for continuous $X$)

# likelihood function

- $\lambda_x : \Theta \to [0, 1]$ is the (relative) likelihood function given $X = x$, when
$$\sup_{\theta \in \Theta} \lambda_x(\theta) = 1 \quad \text{and} \quad \lambda_x(\theta) \propto P_\theta(X = x)$$
(with $\lambda_x(\theta) \propto f_\theta(x)$ as approximation for continuous $X$)

- $\lambda_x$ describes the relative plausibility of the possible values of $\theta$ in the light of the observation $X = x$, and can thus be used as a basis for post-data decision making

# likelihood function

- $\lambda_x : \Theta \to [0,1]$ is the (relative) likelihood function given $X = x$, when
$$\sup_{\theta \in \Theta} \lambda_x(\theta) = 1 \quad \text{and} \quad \lambda_x(\theta) \propto P_\theta(X = x)$$
(with $\lambda_x(\theta) \propto f_\theta(x)$ as approximation for continuous $X$)

- $\lambda_x$ describes the relative plausibility of the possible values of $\theta$ in the light of the observation $X = x$, and can thus be used as a basis for post-data decision making

- prior information can be described by a prior likelihood function: if $X_1$ and $X_2$ are independent, then $\lambda_{(x_1,x_2)} \propto \lambda_{x_1}\,\lambda_{x_2}$ (i.e., when $X_2 = x_2$ is observed, the prior $\lambda_{x_1}$ is updated to the posterior $\lambda_{(x_1,x_2)}$)

# likelihood function

- $\lambda_x : \Theta \to [0, 1]$ is the (relative) likelihood function given $X = x$, when

$$\sup_{\theta \in \Theta} \lambda_x(\theta) = 1 \quad \text{and} \quad \lambda_x(\theta) \propto P_\theta(X = x)$$

(with $\lambda_x(\theta) \propto f_\theta(x)$ as approximation for continuous $X$)

- $\lambda_x$ describes the relative plausibility of the possible values of $\theta$ in the light of the observation $X = x$, and can thus be used as a basis for post-data decision making

- prior information can be described by a prior likelihood function: if $X_1$ and $X_2$ are independent, then $\lambda_{(x_1,x_2)} \propto \lambda_{x_1} \lambda_{x_2}$ (i.e., when $X_2 = x_2$ is observed, the prior $\lambda_{x_1}$ is updated to the posterior $\lambda_{(x_1,x_2)}$)

- strong similarity with the Bayesian approach (both satisfy the likelihood principle): a fundamental advantage of the likelihood approach is the possibility of not using prior information (since $\lambda_{x_1} \equiv 1$ describes complete ignorance)

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

  (i.e., minimize $\int^{\mathsf{S}} W(\,\cdot\,, d)\, \mathrm{d}\Lambda$, the maxitive integral of the loss $W(\,\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

  (i.e., minimize $\int^{S} W(\cdot\,, d)\, \mathrm{d}\Lambda$, the maxitive integral of the loss $W(\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

- e.g., in the previous simple decision example, the MPL decision is "50"

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

  (i.e., minimize $\int^{\mathsf{S}} W(\,\cdot\,, d)\, \mathrm{d}\Lambda$, the maxitive integral of the loss $W(\,\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

- e.g., in the previous simple decision example, the MPL decision is "50"

- point estimation:

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

  (i.e., minimize $\int^{S} W(\,\cdot\,, d)\, \mathrm{d}\Lambda$, the maxitive integral of the loss $W(\,\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

- e.g., in the previous simple decision example, the MPL decision is "50"

- point estimation:
  - $\mathcal{D} = \Theta$ finite

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

  (i.e., minimize $\int^{S} W(\,\cdot\,, d)\, \mathrm{d}\Lambda$, the maxitive integral of the loss $W(\,\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

- e.g., in the previous simple decision example, the MPL decision is "50"

- point estimation:
  - $\mathcal{D} = \Theta$ finite
  - $W(\,\cdot\,, \hat{\theta}) = I_{\Theta \setminus \{\hat{\theta}\}}$ simple loss function

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

  (i.e., minimize $\int^{S} W(\,\cdot\,, d)\, \mathrm{d}\Lambda$, the maxitive integral of the loss $W(\,\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

- e.g., in the previous simple decision example, the MPL decision is "50"

- point estimation:
  - $\mathcal{D} = \Theta$ finite
  - $W(\,\cdot\,, \hat{\theta}) = I_{\Theta \setminus \{\hat{\theta}\}}$ simple loss function
  - MPL decision: **maximum likelihood estimate** $\hat{\theta}_{ML}$

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

  (i.e., minimize $\int^{\mathsf{S}} W(\,\cdot\,, d)\, \mathrm{d}\Lambda$, the maxitive integral of the loss $W(\,\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

- e.g., in the previous simple decision example, the MPL decision is "50"

- point estimation:
  - $\mathcal{D} = \Theta$ finite
  - $W(\,\cdot\,, \hat{\theta}) = I_{\Theta \setminus \{\hat{\theta}\}}$ simple loss function
  - MPL decision: **maximum likelihood estimate** $\hat{\theta}_{ML}$

- hypothesis testing:

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

  (i.e., minimize $\int^{S} W(\,\cdot\,, d)\, \mathrm{d}\Lambda$, the maxitive integral of the loss $W(\,\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

- e.g., in the previous simple decision example, the MPL decision is "50"

- point estimation:
  - $\mathcal{D} = \Theta$ finite
  - $W(\,\cdot\,, \hat{\theta}) = I_{\Theta \setminus \{\hat{\theta}\}}$ simple loss function
  - MPL decision: **maximum likelihood estimate** $\hat{\theta}_{ML}$

- hypothesis testing:
  - $\mathcal{D} = \{H_0, H_1\}$ with $H_0 : \theta \in \mathcal{H}_0 \subset \Theta$ and $H_1 : \theta \in \mathcal{H}_1 = \Theta \setminus \mathcal{H}_0$

# MPL criterion

- MPL criterion: minimize $\sup_{\theta \in \Theta} W(\theta, d)\, \lambda(\theta)$

  (i.e., minimize $\int^{S} W(\,\cdot\,, d)\, \mathrm{d}\Lambda$, the maxitive integral of the loss $W(\,\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta \in \mathcal{H}} \lambda(\theta)$)

- e.g., in the previous simple decision example, the MPL decision is "50"

- point estimation:
  - $\mathcal{D} = \Theta$ finite
  - $W(\,\cdot\,, \hat{\theta}) = I_{\Theta \setminus \{\hat{\theta}\}}$ simple loss function
  - MPL decision: **maximum likelihood estimate** $\hat{\theta}_{ML}$

- hypothesis testing:
  - $\mathcal{D} = \{H_0, H_1\}$ with $H_0 : \theta \in \mathcal{H}_0 \subset \Theta$ and $H_1 : \theta \in \mathcal{H}_1 = \Theta \setminus \mathcal{H}_0$
  - $W(\,\cdot\,, H_1) = c\, I_{\mathcal{H}_0}$ and $W(\,\cdot\,, H_0) = c'\, I_{\mathcal{H}_1}$ with $c \geq c'$

# MPL criterion

- MPL criterion: minimize $\sup_{\theta\in\Theta} W(\theta, d)\,\lambda(\theta)$

  (i.e., minimize $\int^{S} W(\,\cdot\,, d)\,\mathrm{d}\Lambda$, the maxitive integral of the loss $W(\,\cdot\,, d)$ with respect to the maxitive measure $\Lambda : \mathcal{H} \mapsto \sup_{\theta\in\mathcal{H}} \lambda(\theta)$)

- e.g., in the previous simple decision example, the MPL decision is "50"

- point estimation:
  - $\mathcal{D} = \Theta$ finite
  - $W(\,\cdot\,, \hat{\theta}) = I_{\Theta\setminus\{\hat{\theta}\}}$ simple loss function
  - MPL decision: **maximum likelihood estimate** $\hat{\theta}_{ML}$

- hypothesis testing:
  - $\mathcal{D} = \{H_0, H_1\}$ with $H_0 : \theta \in \mathcal{H}_0 \subset \Theta$ and $H_1 : \theta \in \mathcal{H}_1 = \Theta \setminus \mathcal{H}_0$
  - $W(\,\cdot\,, H_1) = c\, I_{\mathcal{H}_0}$ and $W(\,\cdot\,, H_0) = c'\, I_{\mathcal{H}_1}$ with $c \geq c'$
  - MPL decision: **likelihood ratio test** $\Lambda(\mathcal{H}_0) \geq \frac{c'}{c}$

# simple estimation example

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta = (0, +\infty)$ (i.e., $\theta$ positive and $\sigma$ known)

# simple estimation example

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta = (0, +\infty)$ (i.e., $\theta$ positive and $\sigma$ known)

- estimation of $\theta$ with squared error:

# simple estimation example

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta = (0, +\infty)$ (i.e., $\theta$ positive and $\sigma$ known)

- estimation of $\theta$ with squared error:
  - $\mathcal{D} = \Theta$ with $W(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

# simple estimation example

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta = (0, +\infty)$ (i.e., $\theta$ positive and $\sigma$ known)

- estimation of $\theta$ with squared error:
  - $\mathcal{D} = \Theta$ with $W(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
  - no unbiased estimator, maximum likelihood estimator not well-defined, no standard (proper) Bayesian prior

# simple estimation example

- $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta = (0, +\infty)$ (i.e., $\theta$ positive and $\sigma$ known)

- estimation of $\theta$ with squared error:
  - $\mathcal{D} = \Theta$ with $W(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
  - no unbiased estimator, maximum likelihood estimator not well-defined, no standard (proper) Bayesian prior

- estimator resulting from the MPL criterion:

# simple estimation example

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta = (0, +\infty)$ (i.e., $\theta$ positive and $\sigma$ known)

- estimation of $\theta$ with squared error:
  - $\mathcal{D} = \Theta$ with $W(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
  - no unbiased estimator, maximum likelihood estimator not well-defined, no standard (proper) Bayesian prior

- estimator resulting from the MPL criterion:
  - scale invariance and sufficiency: $\hat{\theta}(x_1, \ldots, x_n) = g\left(\frac{\sqrt{n}}{\sigma}\,\bar{x}\right)\frac{\sigma}{\sqrt{n}}$

# simple estimation example

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta = (0, +\infty)$ (i.e., $\theta$ positive and $\sigma$ known)

- estimation of $\theta$ with squared error:
  - $\mathcal{D} = \Theta$ with $W(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
  - no unbiased estimator, maximum likelihood estimator not well-defined, no standard (proper) Bayesian prior

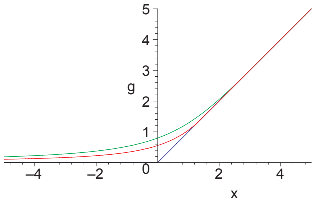
- estimator resulting from the MPL criterion:
  - scale invariance and sufficiency: $\hat{\theta}(x_1, \ldots, x_n) = g\left(\frac{\sqrt{n}}{\sigma}\,\bar{x}\right)\frac{\sigma}{\sqrt{n}}$
  - consistency and asymptotic efficiency: $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$ when $\bar{x} \geq \frac{\sqrt{2}\,\sigma}{\sqrt{n}}$

# simple estimation example

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta = (0, +\infty)$ (i.e., $\theta$ positive and $\sigma$ known)

- estimation of $\theta$ with squared error:
  - $\mathcal{D} = \Theta$ with $W(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
  - no unbiased estimator, maximum likelihood estimator not well-defined, no standard (proper) Bayesian prior

- estimator resulting from the MPL criterion:
  - scale invariance and sufficiency: $\hat{\theta}(x_1, \ldots, x_n) = g\left(\frac{\sqrt{n}}{\sigma}\,\bar{x}\right)\frac{\sigma}{\sqrt{n}}$
  - consistency and asymptotic efficiency: $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$ when $\bar{x} \geq \frac{\sqrt{2}\,\sigma}{\sqrt{n}}$

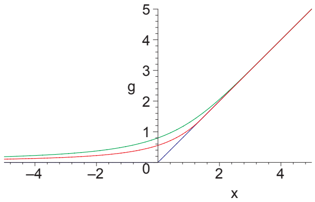


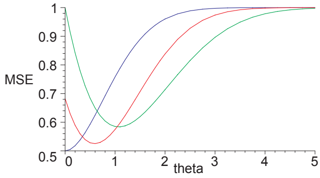
# simple estimation example

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\Theta = (0, +\infty)$ (i.e., $\theta$ positive and $\sigma$ known)

- estimation of $\theta$ with squared error:
  - $\mathcal{D} = \Theta$ with $W(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
  - no unbiased estimator, maximum likelihood estimator not well-defined, no standard (proper) Bayesian prior

- estimator resulting from the MPL criterion:
  - scale invariance and sufficiency: $\hat{\theta}(x_1, \ldots, x_n) = g\left(\frac{\sqrt{n}}{\sigma}\,\bar{x}\right)\frac{\sigma}{\sqrt{n}}$
  - consistency and asymptotic efficiency: $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$ when $\bar{x} \geq \frac{\sqrt{2}\,\sigma}{\sqrt{n}}$



# likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda)$
  (e.g., $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)\, \lambda(\theta)$ for the MPL criterion),

# likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda)$
  (e.g., $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)\, \lambda(\theta)$ for the MPL criterion),
  where the functional $V$ must satisfy the following three properties, for all functions $w, w' : \Theta \to [0, +\infty)$ and all likelihood functions $\lambda, \lambda_n : \Theta \to [0, 1]$

# likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda)$
  (e.g., $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)\, \lambda(\theta)$ for the MPL criterion),
  where the functional $V$ must satisfy the following three properties, for all functions $w, w' : \Theta \to [0, +\infty)$ and all likelihood functions $\lambda, \lambda_n : \Theta \to [0, 1]$
  - monotonicity: $w \leq w'$ (pointwise) $\Rightarrow V(w, \lambda) \leq V(w', \lambda)$

# likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda)$
  (e.g., $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta) \, \lambda(\theta)$ for the MPL criterion),
  where the functional $V$ must satisfy the following three properties, for all functions $w, w' : \Theta \to [0, +\infty)$ and all likelihood functions $\lambda, \lambda_n : \Theta \to [0, 1]$
  - monotonicity: $w \leq w'$ (pointwise) $\Rightarrow V(w, \lambda) \leq V(w', \lambda)$
    (implied by meaning of $W$)

# likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda)$
  (e.g., $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)\, \lambda(\theta)$ for the MPL criterion),
  where the functional $V$ must satisfy the following three properties, for all functions $w, w' : \Theta \to [0, +\infty)$ and all likelihood functions $\lambda, \lambda_n : \Theta \to [0, 1]$
  - monotonicity: $w \leq w'$ (pointwise) $\Rightarrow V(w, \lambda) \leq V(w', \lambda)$
    (implied by meaning of $W$)
  - parametrization invariance: $b : \Theta \to \Theta$ bijection $\Rightarrow V(w \circ b, \lambda \circ b) = V(w, \lambda)$

# likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda)$
  (e.g., $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)\, \lambda(\theta)$ for the MPL criterion),
  where the functional $V$ must satisfy the following three properties, for all functions $w, w' : \Theta \to [0, +\infty)$ and all likelihood functions $\lambda, \lambda_n : \Theta \to [0, 1]$
  - monotonicity: $w \leq w'$ (pointwise) $\Rightarrow V(w, \lambda) \leq V(w', \lambda)$
    (implied by meaning of $W$)
  - parametrization invariance: $b : \Theta \to \Theta$ bijection $\Rightarrow V(w \circ b, \lambda \circ b) = V(w, \lambda)$
    (excludes Bayesian criteria $V(w, \lambda) = \frac{\int w\, \lambda\, d\mu}{\int \lambda\, d\mu}$ for infinite $\Theta$)

# likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda)$
  (e.g., $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)\, \lambda(\theta)$ for the MPL criterion),
  where the functional $V$ must satisfy the following three properties, for all functions $w, w' : \Theta \to [0, +\infty)$ and all likelihood functions $\lambda, \lambda_n : \Theta \to [0, 1]$
  - monotonicity: $w \leq w'$ (pointwise) $\Rightarrow V(w, \lambda) \leq V(w', \lambda)$
    (implied by meaning of $W$)
  - parametrization invariance: $b : \Theta \to \Theta$ bijection $\Rightarrow V(w \circ b, \lambda \circ b) = V(w, \lambda)$
    (excludes Bayesian criteria $V(w, \lambda) = \frac{\int w\, \lambda\, d\mu}{\int \lambda\, d\mu}$ for infinite $\Theta$)
  - consistency: $\mathcal{H} \subseteq \Theta$ with $\lim_{n \to \infty} \sup_{\theta \in \Theta \setminus \mathcal{H}} \lambda_n(\theta) = 0 \Rightarrow$
    $\lim_{n \to \infty} V(c\, I_{\mathcal{H}} + c'\, I_{\Theta \setminus \mathcal{H}}, \lambda_n) = c$ for all constants $c, c' \in [0, +\infty)$

# likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda)$
  (e.g., $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)\,\lambda(\theta)$ for the MPL criterion),
  where the functional $V$ must satisfy the following three properties, for all functions $w, w' : \Theta \to [0, +\infty)$ and all likelihood functions $\lambda, \lambda_n : \Theta \to [0, 1]$
  - monotonicity: $w \leq w'$ (pointwise) $\Rightarrow V(w, \lambda) \leq V(w', \lambda)$
    (implied by meaning of $W$)
  - parametrization invariance: $b : \Theta \to \Theta$ bijection $\Rightarrow V(w \circ b, \lambda \circ b) = V(w, \lambda)$
    (excludes Bayesian criteria $V(w, \lambda) = \frac{\int w\,\lambda\, d\mu}{\int \lambda\, d\mu}$ for infinite $\Theta$)
  - consistency: $\mathcal{H} \subseteq \Theta$ with $\lim_{n \to \infty} \sup_{\theta \in \Theta \setminus \mathcal{H}} \lambda_n(\theta) = 0 \Rightarrow$
    $\lim_{n \to \infty} V(c\, I_{\mathcal{H}} + c'\, I_{\Theta \setminus \mathcal{H}},\, \lambda_n) = c$ for all constants $c, c' \in [0, +\infty)$
    (excludes minimax criterion $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)$,
    implies calibration: $V(c, \lambda) = c$)

# likelihood decision criteria

- likelihood decision criterion: minimize $V(W(\cdot, d), \lambda)$
  (e.g., $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)\,\lambda(\theta)$ for the MPL criterion),
  where the functional $V$ must satisfy the following three properties, for all functions $w, w' : \Theta \to [0, +\infty)$ and all likelihood functions $\lambda, \lambda_n : \Theta \to [0, 1]$
  - monotonicity: $w \leq w'$ (pointwise) $\Rightarrow V(w, \lambda) \leq V(w', \lambda)$
    (implied by meaning of $W$)
  - parametrization invariance: $b : \Theta \to \Theta$ bijection $\Rightarrow V(w \circ b, \lambda \circ b) = V(w, \lambda)$
    (excludes Bayesian criteria $V(w, \lambda) = \frac{\int w\,\lambda\,d\mu}{\int \lambda\,d\mu}$ for infinite $\Theta$)
  - consistency: $\mathcal{H} \subseteq \Theta$ with $\lim_{n\to\infty} \sup_{\theta \in \Theta \setminus \mathcal{H}} \lambda_n(\theta) = 0 \Rightarrow$
    $\lim_{n\to\infty} V(c\, I_{\mathcal{H}} + c'\, I_{\Theta \setminus \mathcal{H}}, \lambda_n) = c$ for all constants $c, c' \in [0, +\infty)$
    (excludes minimax criterion $V(w, \lambda) = \sup_{\theta \in \Theta} w(\theta)$,
    implies calibration: $V(c, \lambda) = c$)

- **likelihood decision function**: $\delta : \mathcal{X} \to \mathcal{D}$ such that $\delta(x)$ minimizes $V(W(\cdot, d), \lambda_x)$ for all $x \in \mathcal{X}$

# properties

- likelihood decision criteria have the advantages of **post-data** methods:

# properties

- likelihood decision criteria have the advantages of **post-data** methods:
  - independence from choice of possible alternative observations

# properties

- likelihood decision criteria have the advantages of **post-data** methods:
  - independence from choice of possible alternative observations
  - direct interpretation

# properties

- likelihood decision criteria have the advantages of **post-data** methods:
  - independence from choice of possible alternative observations
  - direct interpretation
  - simpler problems

# properties

- likelihood decision criteria have the advantages of **post-data** methods:
  - independence from choice of possible alternative observations
  - direct interpretation
  - simpler problems
- likelihood decision criteria have also important **pre-data** properties:

# properties

- likelihood decision criteria have the advantages of **post-data** methods:
  - independence from choice of possible alternative observations
  - direct interpretation
  - simpler problems
- likelihood decision criteria have also important **pre-data** properties:
  - equivariance: for invariant decision problems, the likelihood decision functions are equivariant

# properties

- likelihood decision criteria have the advantages of **post-data** methods:
  - independence from choice of possible alternative observations
  - direct interpretation
  - simpler problems
- likelihood decision criteria have also important **pre-data** properties:
  - equivariance: for invariant decision problems, the likelihood decision functions are equivariant
  - consistency: under some regularity conditions, the likelihood decision functions $\delta_n : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathcal{D}$ satisfy

$$\lim_{n \to \infty} W(\theta, \delta_n(X_1, \ldots, X_n)) = \inf_{d \in \mathcal{D}} W(\theta, d) \quad P_\theta\text{-a.s.}$$

# properties

- likelihood decision criteria have the advantages of **post-data** methods:
  - independence from choice of possible alternative observations
  - direct interpretation
  - simpler problems
- likelihood decision criteria have also important **pre-data** properties:
  - equivariance: for invariant decision problems, the likelihood decision functions are equivariant
  - consistency: under some regularity conditions, the likelihood decision functions $\delta_n : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathcal{D}$ satisfy

$$\lim_{n \to \infty} W(\theta, \delta_n(X_1, \ldots, X_n)) = \inf_{d \in \mathcal{D}} W(\theta, d) \quad P_\theta\text{-a.s.}$$

  - asymptotic efficiency: under slightly stronger regularity conditions, the above convergence is as fast as possible

# example: estimation of variance components

- estimation of the variance components in the $3 \times 3$ random effect one-way layout, under normality assumptions and weighted squared error loss

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{for all } i, j \in \{1, 2, 3\}$$

# example: estimation of variance components

- estimation of the variance components in the $3 \times 3$ random effect one-way layout, under normality assumptions and weighted squared error loss

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{for all } i, j \in \{1, 2, 3\}$$

- normality assumptions:

$$\alpha_i \sim \mathcal{N}(0, v_a), \;\; \varepsilon_{ij} \sim \mathcal{N}(0, v_e), \;\; \text{all independent}$$

$$\Rightarrow \;\; X_{ij} \sim \mathcal{N}(\mu,\, v_a + v_e) \;\text{ dependent}, \;\; \theta = (\mu, v_a, v_e) \in \mathbb{R} \times (0, \infty)^2$$

# example: estimation of variance components

- estimates $\widehat{v_e}$ and $\widehat{v_a}$ of variance components $v_e$ and $v_a$ are functions of

$$SS_e = \sum_{i=1}^{3} \sum_{j=1}^{3} (x_{ij} - \bar{x}_{i.})^2 \quad \text{and} \quad SS_a = 3 \sum_{i=1}^{3} (\bar{x}_{i.} - \bar{x}_{..})^2,$$

where

$$\bar{x}_{i.} = \frac{1}{3} \sum_{j=1}^{3} x_{ij}, \quad \bar{x}_{..} = \frac{1}{9} \sum_{i=1}^{3} \sum_{j=1}^{3} x_{ij},$$

$$\frac{SS_e}{v_e} \sim \chi^2_6, \quad \text{and} \quad \frac{\frac{1}{3}\, SS_a}{v_a + \frac{1}{3}\, v_e} \sim \chi^2_2$$

# example: estimation of variance components

- estimates $\widehat{v_e}$ and $\widehat{v_a}$ of variance components $v_e$ and $v_a$ are functions of

$$SS_e = \sum_{i=1}^{3} \sum_{j=1}^{3} (x_{ij} - \bar{x}_{i\cdot})^2 \quad \text{and} \quad SS_a = 3 \sum_{i=1}^{3} (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2,$$

  where

$$\bar{x}_{i\cdot} = \frac{1}{3} \sum_{j=1}^{3} x_{ij}, \quad \bar{x}_{\cdot\cdot} = \frac{1}{9} \sum_{i=1}^{3} \sum_{j=1}^{3} x_{ij},$$

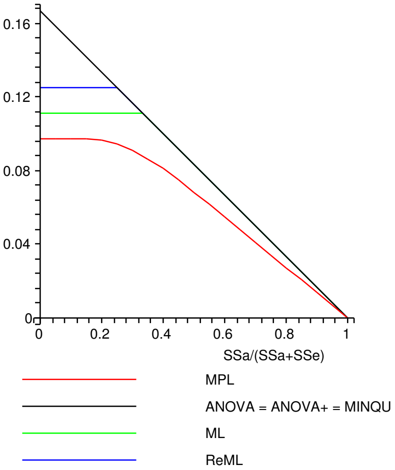
$$\frac{SS_e}{v_e} \sim \chi^2_6, \quad \text{and} \quad \frac{\frac{1}{3}\, SS_a}{v_a + \frac{1}{3}\, v_e} \sim \chi^2_2$$

- invariant loss functions:

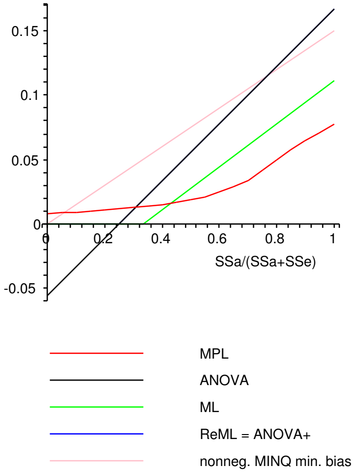
$$W(\theta, \widehat{v_e}) = 3\, \frac{(v_e - \widehat{v_e})^2}{{v_e}^2} \quad \text{and} \quad W(\theta, \widehat{v_a}) = \frac{(v_a - \widehat{v_a})^2}{(v_a + \frac{1}{3}\, v_e)^2}$$

# example: estimation of variance components

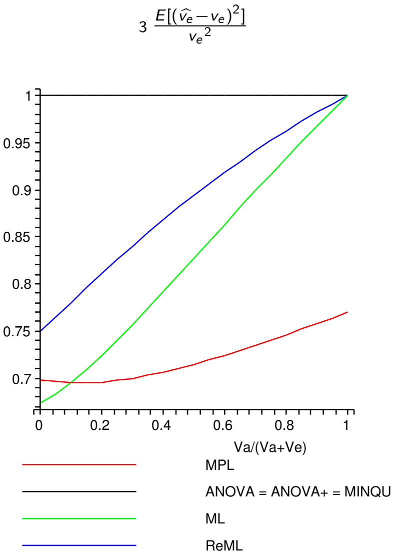
$$\frac{\widehat{v_e}}{SS_a+SS_e}$$

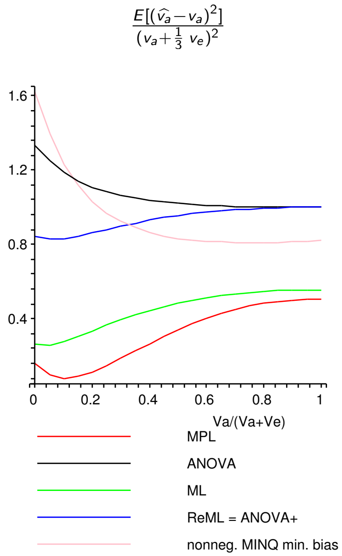



$$\frac{\widehat{v_a}}{SS_a+SS_e}$$



# example: estimation of variance components



# conclusion

- this work:
  - fills a gap in the likelihood approach to statistics

# conclusion

- this work:
  - fills a gap in the likelihood approach to statistics
  - introduces an alternative to classical and Bayesian decision making

# conclusion

- this work:
  - fills a gap in the likelihood approach to statistics
  - introduces an alternative to classical and Bayesian decision making
  - offers a new perspective on the likelihood methods

# conclusion

- this work:
  - fills a gap in the likelihood approach to statistics
  - introduces an alternative to classical and Bayesian decision making
  - offers a new perspective on the likelihood methods
- **likelihood decision making**:
  - is post-data and equivariant

# conclusion

- this work:
  - fills a gap in the likelihood approach to statistics
  - introduces an alternative to classical and Bayesian decision making
  - offers a new perspective on the likelihood methods
- **likelihood decision making**:
  - is post-data and equivariant
  - is consistent and asymptotically efficient

# conclusion

- this work:
  - fills a gap in the likelihood approach to statistics
  - introduces an alternative to classical and Bayesian decision making
  - offers a new perspective on the likelihood methods
- **likelihood decision making**:
  - is post-data and equivariant
  - is consistent and asymptotically efficient
  - does not need prior information

## references


- Lehmann (1959). **Testing Statistical Hypotheses**. Wiley.
- Diehl and Sprott (1965). **Die Likelihoodfunktion und ihre Verwendung beim statistischen Schluß**. *Statistische Hefte* 6, 112–134.
- Giang and Shenoy (2005). **Decision making on the sole basis of statistical likelihood**. *Artificial Intelligence* 165, 137–163.
- Cattaneo (2013). **Likelihood decision functions**. *Electronic Journal of Statistics* 7, 2924–2946.
- Cattaneo (2013). **On maxitive integration**. Technical Report 147, Department of Statistics, LMU Munich.
- Cattaneo and Wiencierz (2012). **Likelihood-based Imprecise Regression**. *International Journal of Approximate Reasoning* 53, 1137–1154.
- Antonucci, Cattaneo, and Corani (2012). **Likelihood-based robust classification with Bayesian networks**. In: *Advances in Computational Intelligence, Part 3*, Springer, 491–500.