Statistical inference: Decision-theoretic perspective

Robert Schlicht

TU Dresden
Fakultät Umweltwissenschaften

2015-10-08

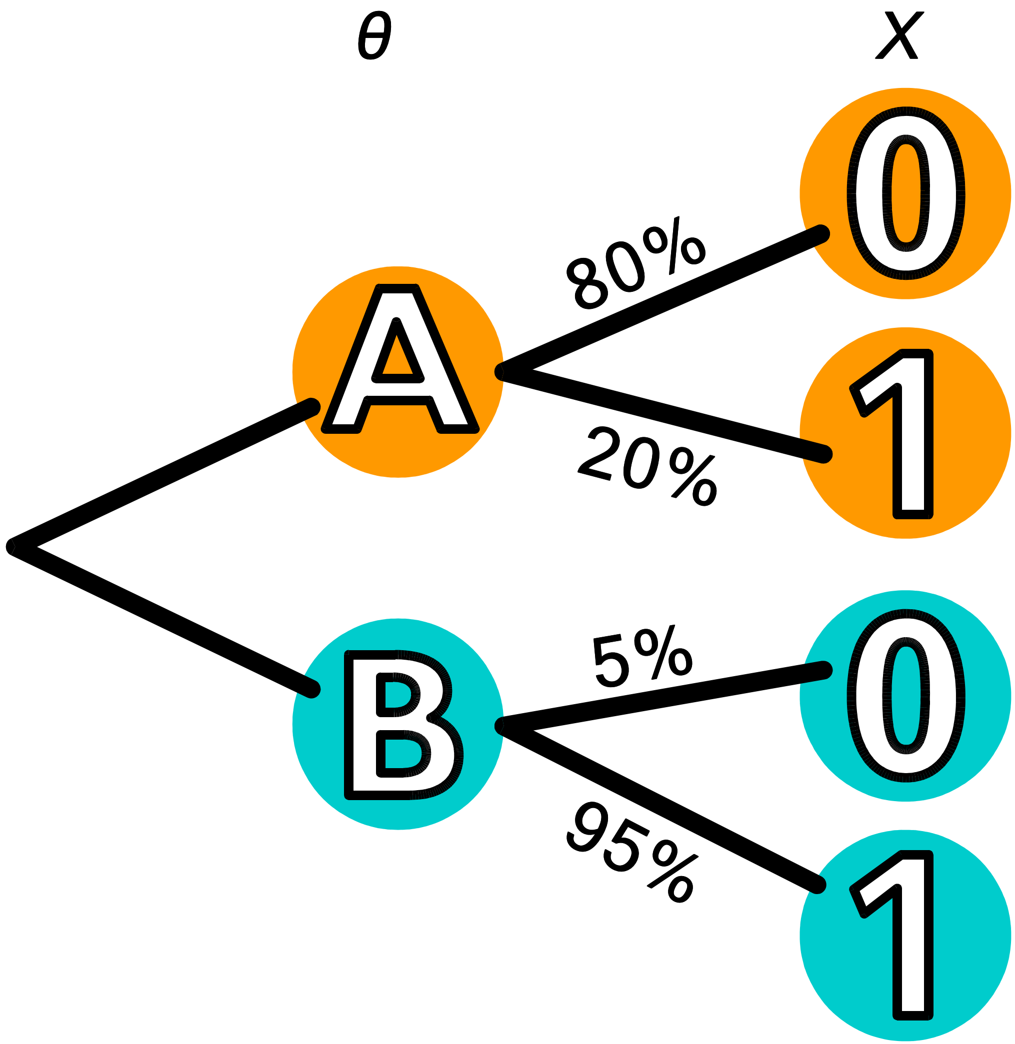
IBS-DR Biometry Workshop

Würzburg University

Quercus petraea or *Quercus robur*?

Quercus petraea or *Quercus robur*?

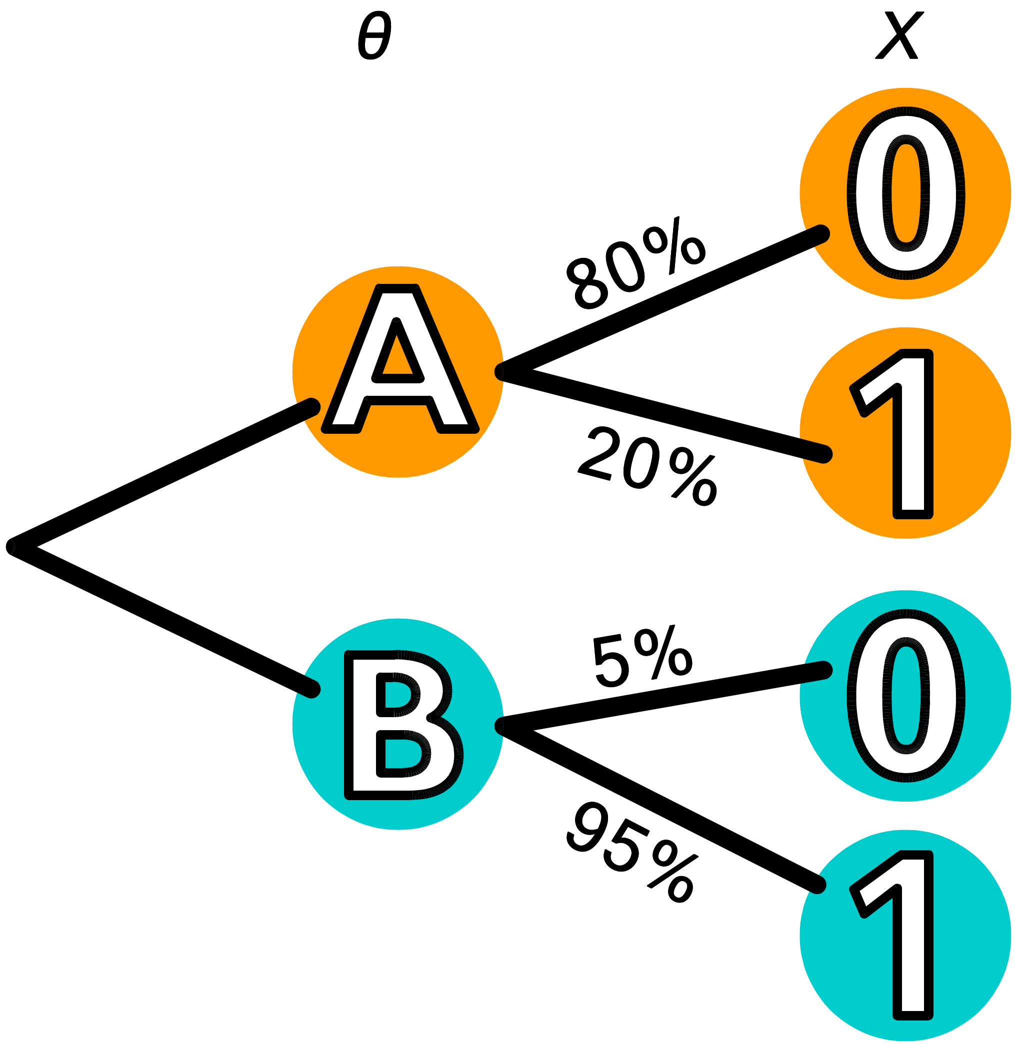
- parameter space {A, B}, sample space {0, 1} (sample X of size 1)



- unknown param. θ = species
  - A: *Quercus petraea*
  - B: *Quercus robur*
- obs. X = length of acorn stalk
  - 0: no long stalk visible
  - 1: long stalk visible

Quercus petraea or *Quercus robur*?

- parameter space {A, B}, sample space {0, 1} (sample X of size 1)



- estimator for θ:

$$\hat{\theta} = \begin{cases} A & \text{if } X = 0 \\ B & \text{if } X = 1 \end{cases}$$

(maximum likelihood)

- 95%-confidence region for θ:

$$C = \begin{cases} \{A\} & \text{if } X = 0 \\ \{A, B\} & \text{if } X = 1 \end{cases}$$

- 5%-test for $H_0 : \theta = B$: Reject H_0 if and only if $X = 0$. (power = 80%)

Statistical decision theory

Basic approach (classical): (Fisher 1921, Neyman & Pearson 1933)

(1) Choose a strategy that, *before* the observation, leads to reasonable results with high probability for every possible parameter value.

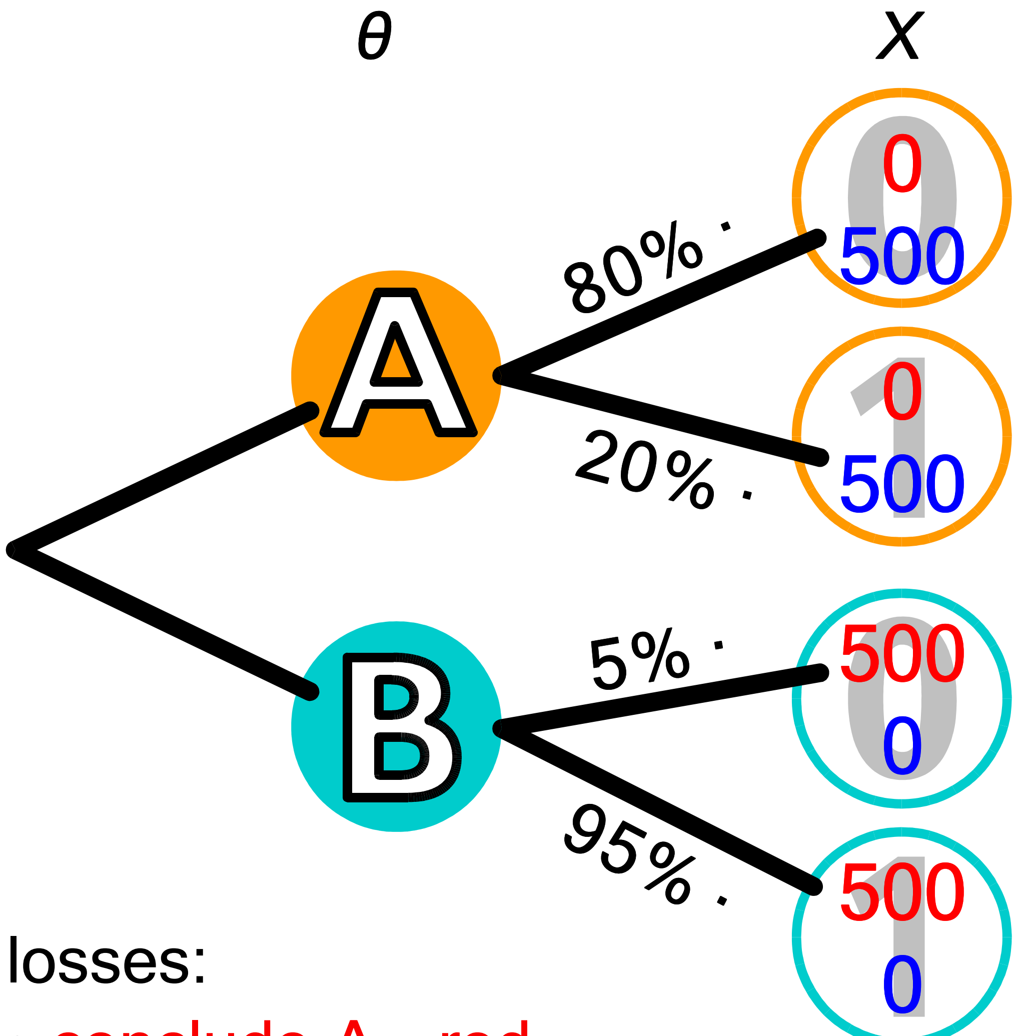
(2) Stick to that strategy *after* the observation.

Broad framework: (Wald 1940s)

- Specify the loss associated with every possible action as a function of θ (loss function).
- Compare different decision rules by looking at the expected value (mean) of the loss, as a function of θ (risk function).

Classical approach: Estimating θ

- loss for estimator $\hat{\theta}$: 0 if $\hat{\theta} = \theta$ and 500 if $\hat{\theta} \neq \theta$

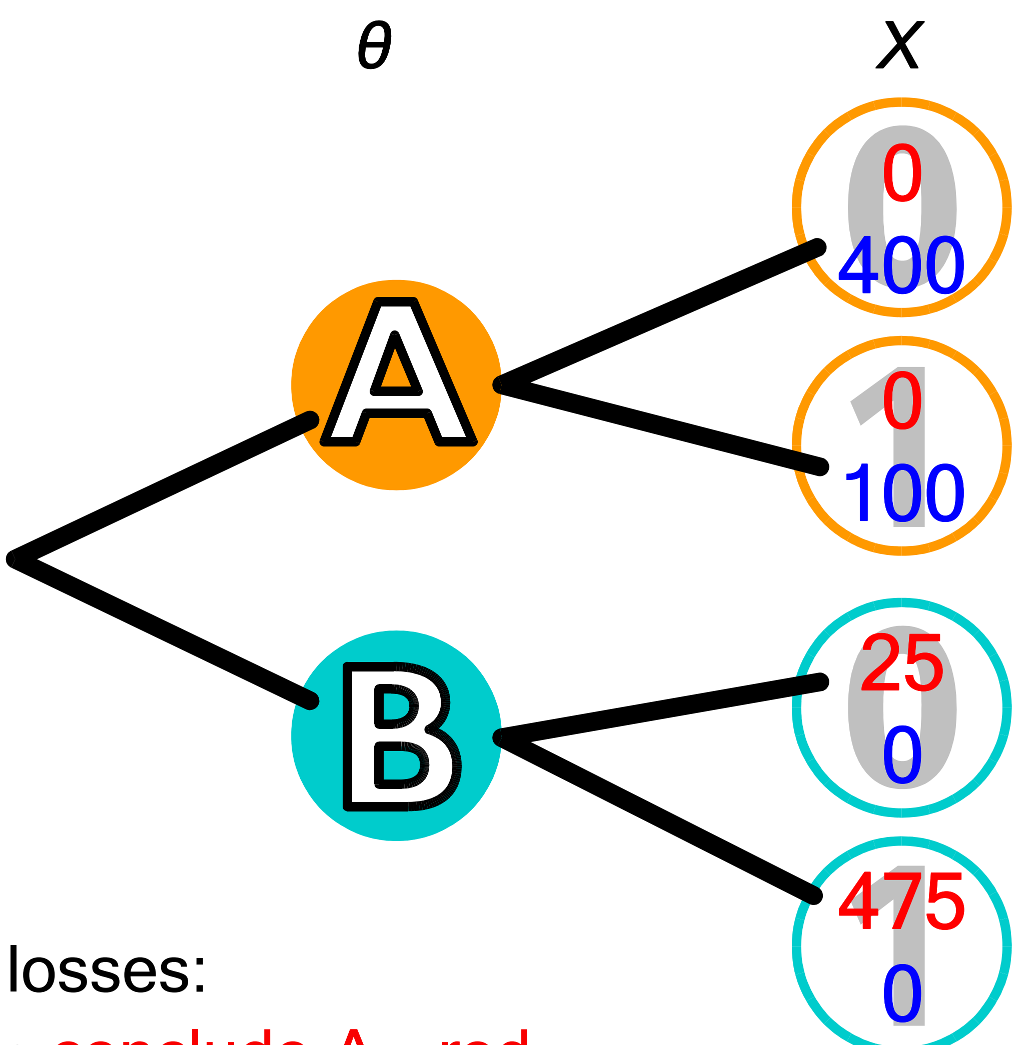


losses:

- conclude A - red
- conclude B - blue

Classical approach: Estimating θ

- loss for estimator $\hat{\theta}$: 0 if $\hat{\theta} = \theta$ and 500 if $\hat{\theta} \neq \theta$



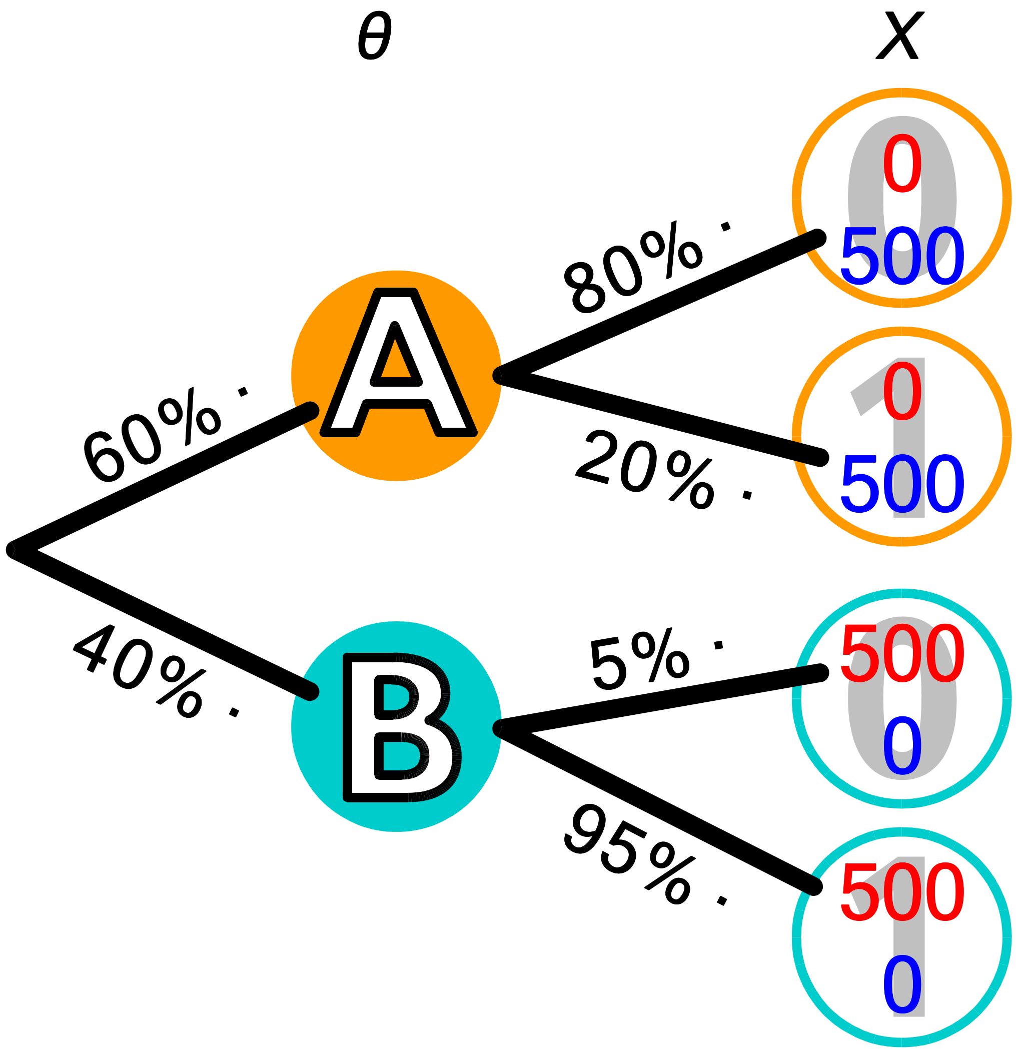
estimator	mean loss if θ = A	mean loss if θ = B
$\hat{\theta}_1$ = A	0	500
$\hat{\theta}_2 = \begin{cases} A & \text{if } X=0 \\ B & \text{if } X=1 \end{cases}$	100	25
$\hat{\theta}_3 = \begin{cases} B & \text{if } X=0 \\ A & \text{if } X=1 \end{cases}$	400	475
$\hat{\theta}_4$ = B	500	0

- $\hat{\theta}_3$ is *not admissible* ($\hat{\theta}_2$ better).
- $\hat{\theta}_2$ is *minimax* estimator (minimizes the maximum mean loss).

losses:
- conclude A - red
- conclude B - blue

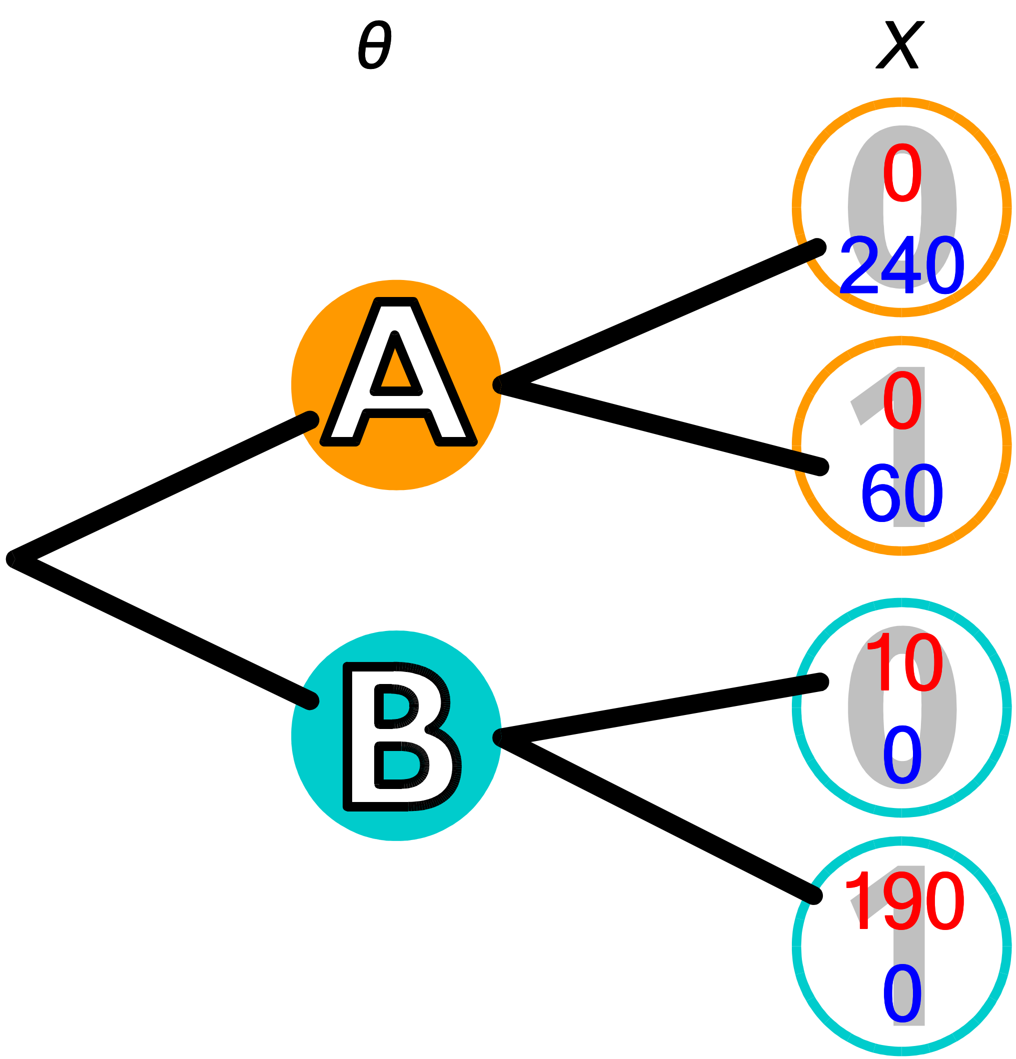
Bayesian approach: Estimating θ

- loss for estimator $\hat{\theta}$: 0 if $\hat{\theta} = \theta$ and 500 if $\hat{\theta} \neq \theta$



Bayesian approach: Estimating θ

- loss for estimator $\hat{\theta}$: 0 if $\hat{\theta} = \theta$ and 500 if $\hat{\theta} \neq \theta$



- Conclusions can be based on non-conditional mean losses:

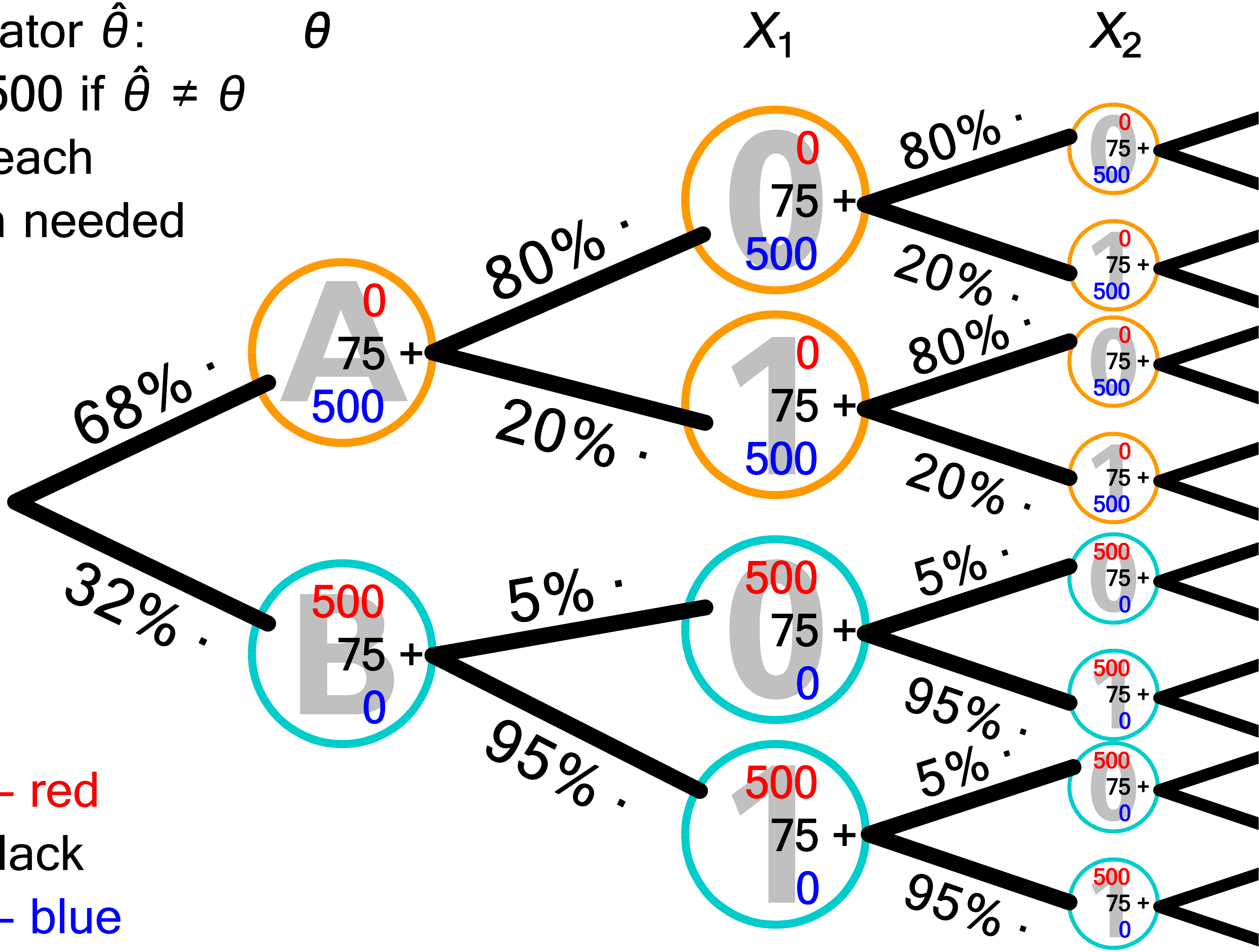
estimator	*mean loss*
$\hat{\theta}_1$ = A	200
$\hat{\theta}_2 = \begin{cases} A & \text{if } X = 0 \\ B & \text{if } X = 1 \end{cases}$	70
$\hat{\theta}_3 = \begin{cases} B & \text{if } X = 0 \\ A & \text{if } X = 1 \end{cases}$	430
$\hat{\theta}_4$ = B	300

- $\hat{\theta}_2$ is the unique optimal Bayes estimator.

Sequential procedure (Bayesian): Estimating θ

loss for estimator $\hat{\theta}$:

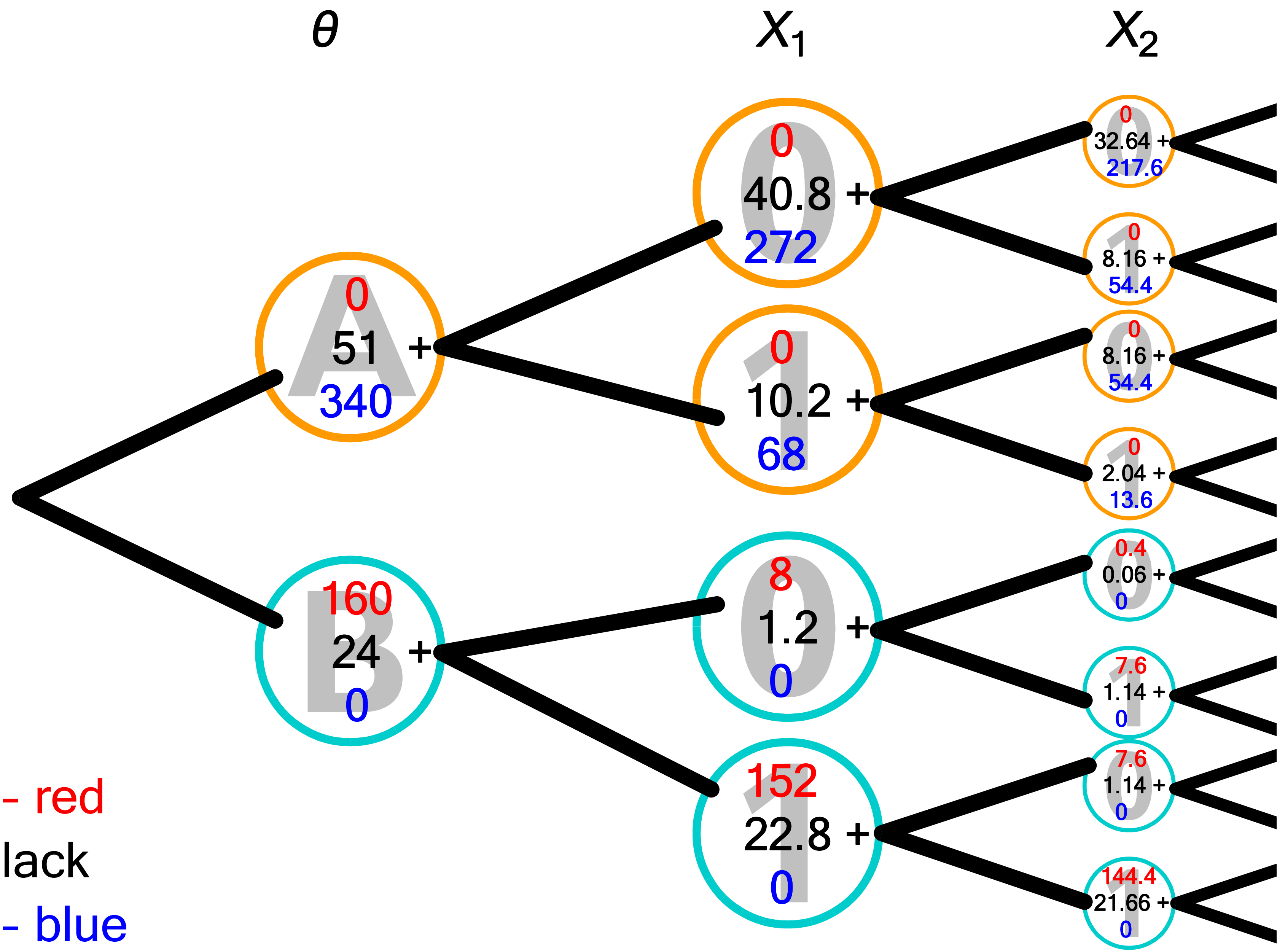
- 0 if $\hat{\theta} = \theta$, 500 if $\hat{\theta} \neq \theta$
- plus 75 for each observation needed



losses:

- conclude A - red
- continue - black
- conclude B - blue

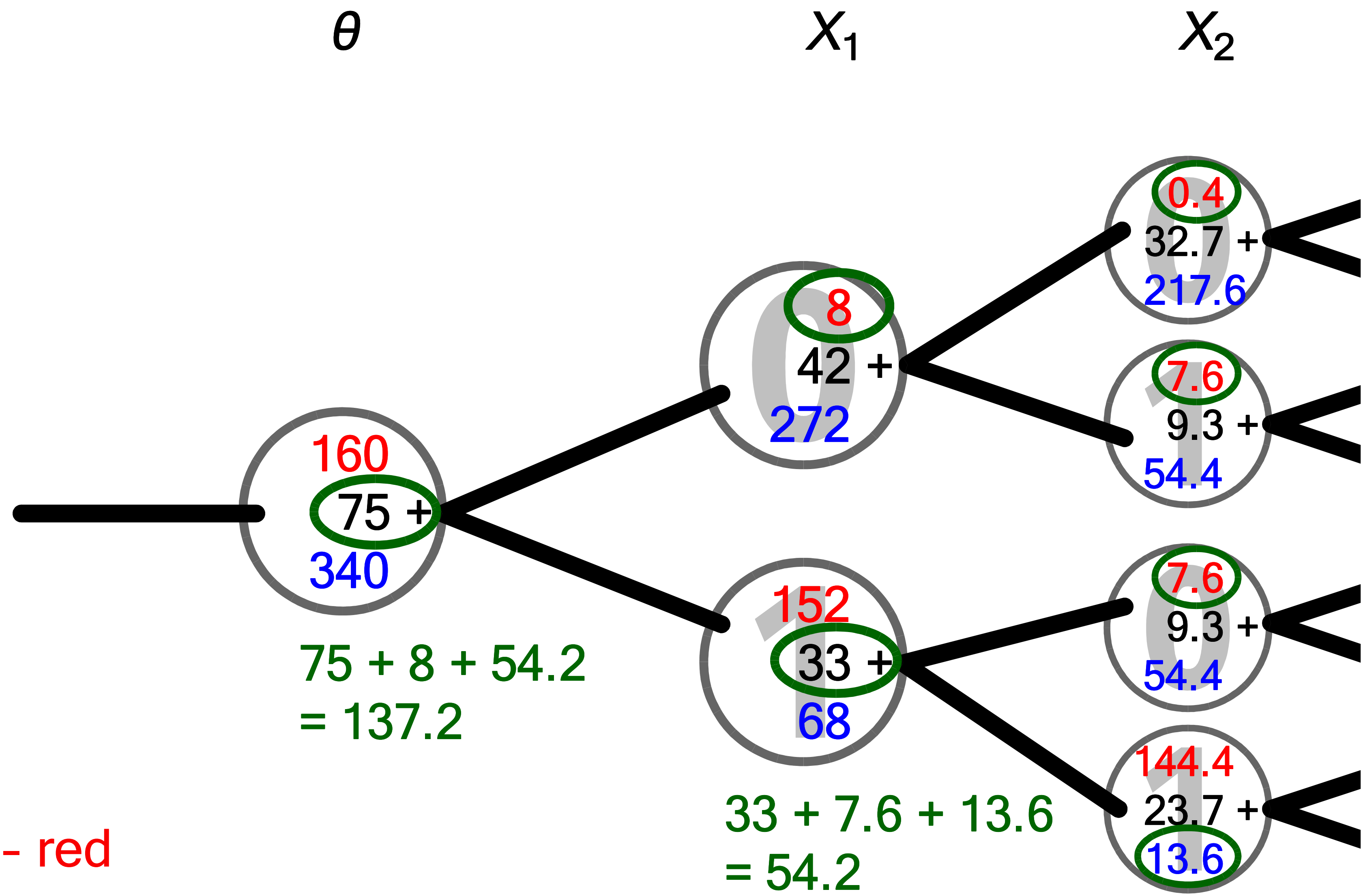
Sequential procedure (Bayesian): Estimating θ



losses:

- conclude A - red
- continue - black
- conclude B - blue

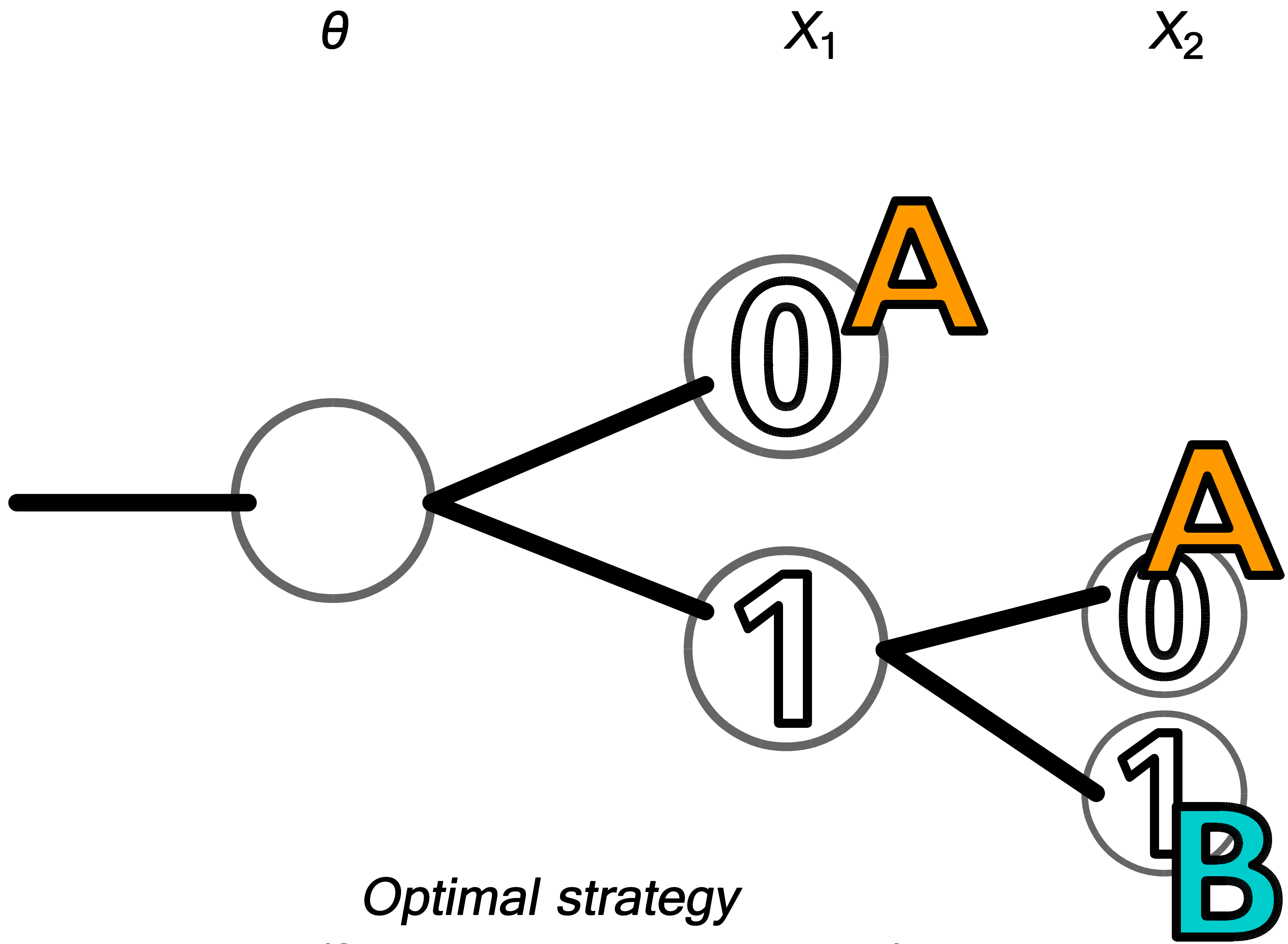
Sequential procedure (Bayesian): Estimating θ



losses:

- conclude A - red
- continue - black
- conclude B - blue

Sequential procedure (Bayesian): Estimating θ



Optimal strategy
(for given prior and losses)

Problems

- Classical or Bayesian – which one is better?

- Classical approach: among admissible decision rules (minimax, minimax-regret, ...), which choice is the best?

- Can we justify basing inference on probabilities or expected values? What *is* probability?

Probability: degree an event is supposed at to occur

- maximum value 100% = 1 for events to be treated like sure ones
- (countably) additive

Probability: degree an event is supposed at to occur

- maximum value 100% = 1 for events to be treated like sure ones
- (countably) additive (Kolmogorov 1933)

Frequentist interpretation:

a. (random sampling:) *probability = proportion of population* when drawing an element at random (all equally probable)

b. (random experiments:) *probability = long-term relative frequency* almost surely for independent repetitions under uniform conditions *(law of large numbers)* (J. Bernoulli 1713, Kolmogorov, de Finetti)

Probability: degree an event is supposed at to occur

- maximum value 100% = 1 for events to be treated like sure ones
- (countably) additive (Kolmogorov 1933)

Frequentist interpretation:

a. (random sampling:) *probability = proportion of population* when drawing an element at random (all equally probable)

b. (random experiments:) *probability = long-term relative frequency* almost surely for independent repetitions under uniform conditions *(law of large numbers)* (J. Bernoulli 1713, Kolmogorov, de Finetti)

⇒ Other interpretations usually lead to the same probabilities (consider a hypothetical population or random experiment).

Probabilities can model degrees of belief - **subjective interpretation.**

Probabilities manifest themselves in decisions. (Ramsey 1926, de Finetti)

Decision-theoretic concept of probability

Let $(S, \mathscr{S})$ be totally bounded space. If and only if S is complete, the decision principle

$$f \succcurlyeq g \quad \Leftrightarrow \quad \text{expect. of } f \geq \text{expect. of } g \text{ for all measures in } C^*$$

defines a one-to-one correspondence between

- closed convex sets C^* of regular probability measures on S,
- preference relations $\succcurlyeq$ on $\mathscr{S}$

Decision-theoretic concept of probability

Let $(S, \mathscr{S})$ be totally bounded space. If and only if S is complete, the decision principle

$$f \succcurlyeq g \quad \Leftrightarrow \quad \text{expect. of } f \geq \text{expect. of } g \text{ for all measures in } C^*$$

defines a one-to-one correspondence between

- closed convex sets C^* of regular probability measures on S,
- preference relations $\succcurlyeq$ on $\mathscr{S}$

Probabilities are here the means of deriving decisions consistent with a given initial set of decisions.

Totally bounded spaces

A **totally bounded space** is a set S together with a linear space $\mathscr{S}$ of bounded real-valued functions on S that includes all constant functions and is closed under uniform limits and pointwise maxima (or multiplication).

Totally bounded spaces

A **totally bounded space** is a set S together with a linear space $\mathcal{S}$ of bounded real-valued functions on S that includes all constant functions and is closed under uniform limits and pointwise maxima (or multiplication).

In S we have the topology generated by the functions in $\mathcal{S}$, i. e. x_n converges to x if and only if $\lim f(x_n) = f(x)$ for all f in $\mathcal{S}$.

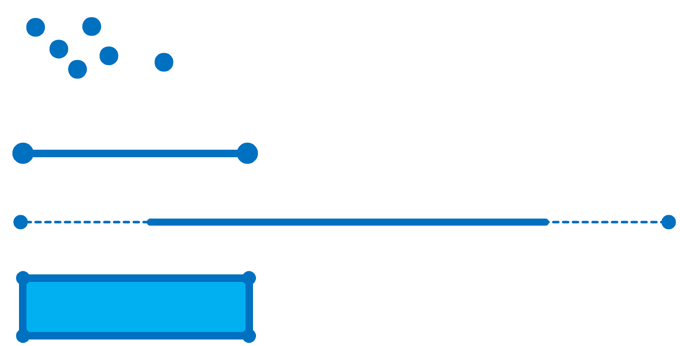
Totally bounded spaces

A **totally bounded space** is a set S together with a linear space $\mathscr{S}$ of bounded real-valued functions on S that includes all constant functions and is closed under uniform limits and pointwise maxima (or multiplication).

In S we have the topology generated by the functions in $\mathscr{S}$, i. e. x_n converges to x if and only if $\lim f(x_n) = f(x)$ for all f in $\mathscr{S}$.

S is **complete** if existence of $\lim f(x_n)$ for all f in $\mathscr{S}$ implies convergence to an x in S, for every net (x_n). *These are precisely the compact regular topological spaces S with the set $\mathscr{S}$ of continuous real-valued functions. Examples:*

- all finite sets
- all closed intervals
- the extended real numbers
- any products of such spaces



Regular probability measures

A **regular probability measure** on a totally bounded space $(S, \mathscr{S})$ is a normalized positive linear functional m on $\mathscr{S}$ that can be extended to a set including indicator functions of closed sets such that for all φ, ψ into totally bounded spaces $(S', \mathscr{S}')$ and factors α, whenever $f \mapsto m(f \circ \varphi) - \alpha\, m(f \circ \psi)$, $f \in \mathscr{S}'$, is a positive linear functional, the same is true for the extension.

Regular probability measures

A **regular probability measure** on a totally bounded space $(S, \mathscr{S})$ is a normalized positive linear functional m on $\mathscr{S}$ that can be extended to a set including indicator functions of closed sets such that for all φ, ψ into totally bounded spaces $(S', \mathscr{S}')$ and factors α, whenever $f \mapsto m(f \circ \varphi) - \alpha\, m(f \circ \psi)$, $f \in \mathscr{S}'$, is a positive linear functional, the same is true for the extension.

Equivalently, m defines an integral $\int f(x)\, m(dx)$ ("expected value of f") equal to $\lim m(f_n)$ *for limits of increasing (or decreasing) nets (f_n) in $\mathscr{S}$ and with the usual properties for Borel measurable functions.*

Regular probability measures

A **regular probability measure** on a totally bounded space $(S, \mathscr{S})$ is a normalized positive linear functional m on $\mathscr{S}$ that can be extended to a set including indicator functions of closed sets such that for all φ, ψ into totally bounded spaces $(S', \mathscr{S}')$ and factors α, whenever $f \mapsto m(f \circ \varphi) - \alpha\, m(f \circ \psi)$, $f \in \mathscr{S}'$, is a positive linear functional, the same is true for the extension.

Equivalently, m defines an integral $\int f(x)\, m(dx)$ ("expected value of f") equal to $\lim m(f_n)$ *for limits of increasing (or decreasing) nets (f_n) in $\mathscr{S}$ and with the usual properties for Borel measurable functions.*

Example: If S is a separable metrizable space (finite set, interval, ...), these are precisely the integrals for Borel probability measures.

Convergence $m_n \to m$ is again defined by $\lim m_n(f) = m(f)$ for all f in $\mathscr{S}$.

Theorem

Let (S, $\mathscr{S}$) be totally bounded space. If and only if S is complete, the decision principle

$$f \succcurlyeq g \quad \Leftrightarrow \quad m(f) \geq m(g) \text{ for all } m \text{ in } C^*$$

defines a one-to-one correspondence between

- closed convex sets C^* of regular probability measures on S,
- relations $\succcurlyeq$ on $\mathscr{S}$ with the following properties:
  (1) $f \succcurlyeq g$ depends only on the difference $f - g$,
  (2) $f \succcurlyeq g$ is implied by each of $f \geq g$; $\alpha f \succcurlyeq \alpha g$ for some $\alpha > 0$;
  $f \succcurlyeq 0 \succcurlyeq g$; $f + c \succcurlyeq g$ for all sufficiently small $c > 0$.

(Schlicht 2015 *Positivity*; cf. Walley 1991, Statist. Reasoning with Imprecise Prob.)

Theorem

Let $(S, \mathscr{S})$ be totally bounded space. If and only if S is complete, the decision principle

$$f \succcurlyeq g \quad \Leftrightarrow \quad m(f) \geq m(g) \text{ for all } m \text{ in } C^*$$

defines a one-to-one correspondence between

- closed convex sets C^* of regular probability measures on S,
- relations $\succcurlyeq$ on $\mathscr{S}$ with the following properties:
  (1) $f \succcurlyeq g$ depends only on the difference $f - g$,
  (2) $f \succcurlyeq g$ is implied by each of $f \geq g$; $\alpha f \succcurlyeq \alpha g$ for some $\alpha > 0$;
  $f \succcurlyeq 0 \succcurlyeq g$; $f + c \succcurlyeq g$ for all sufficiently small $c > 0$.

It is sufficient to assume $\succcurlyeq$ is defined only on those functions in $\mathscr{S}$ with values in a given neighborhood of 0; this can help justifying (1) and (2).

Theorem

The theorem gives a decision-theoretic justification for

- classical statistical models: a set C^* corresponding to all possible priors,
- Bayesian models: a set C^* with a single element,
- between these two extremes: a wide range of models where prior information is only partially probabilistically determined:
  - robust Bayesian approaches
  - imprecise probability models
  - many others

Example: Mixed-effect models

Responses in a linear mixed model are linear combinations of
(1) completely unknown coefficients that determine “fixed effects”,
(2) coefficients following distributions that determine “random effects”,
(3) stochastic terms (“errors”).

The coefficients in (2) differ from the terms in (3) in that their values are a target of statistical inference.

Here we have a situation in which the unknown parameters are partly determined by probabilistic prior information.

Example: Prediction

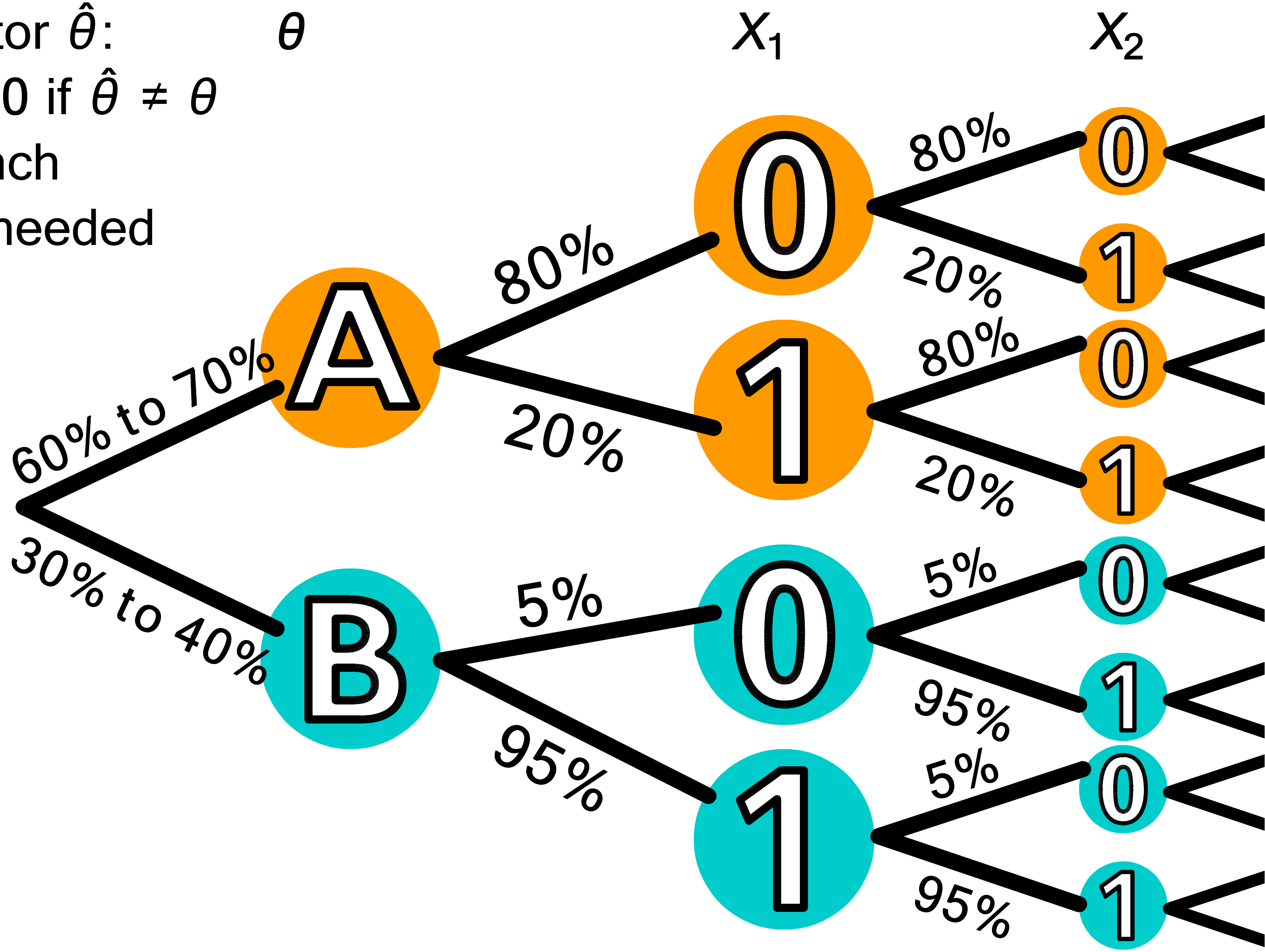
Often we want to draw conclusions not only about unknown parameters, but also about unknown future observations (prediction).

So again we have a situation in which quantities with partially determined probability distributions are the target of statistical inference.

Example: Sequential estimation of θ

loss for estimator $\hat{\theta}$:

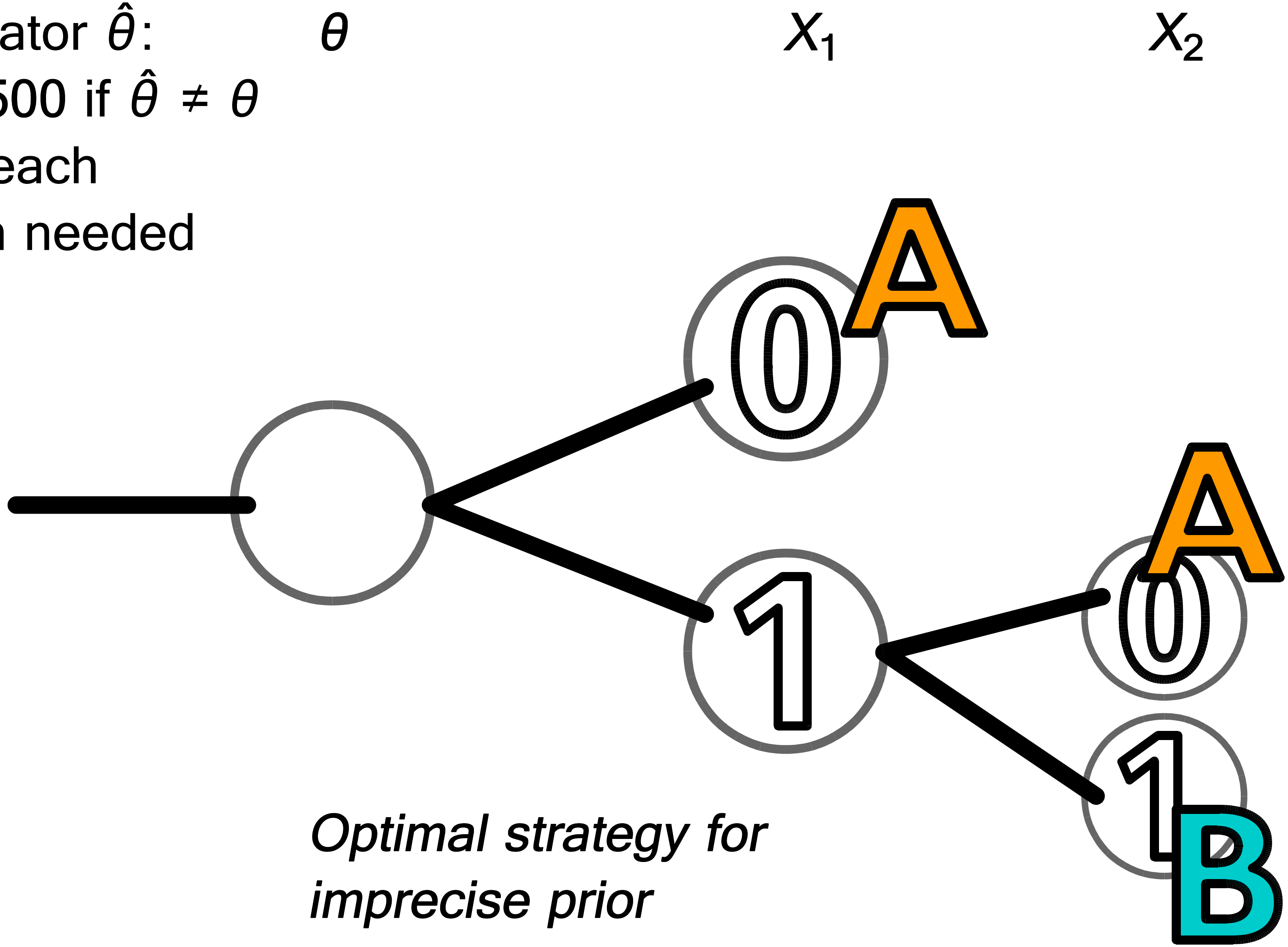
- 0 if $\hat{\theta} = \theta$, 500 if $\hat{\theta} \neq \theta$
- plus 75 for each observation needed



Example: Sequential estimation of θ

loss for estimator $\hat{\theta}$:

- 0 if $\hat{\theta} = \theta$, 500 if $\hat{\theta} \neq \theta$
- plus 75 for each observation needed



Optimal strategy for imprecise prior

$30\% \leq P(\theta = B) \leq 40\%$

How to decide in case there is no single optimal strategy

Solution 1 (honest): *If we cannot decide, we cannot decide.*

- There are situations with too little information, or too much information that cannot be classified, to make well-founded decisions.
- Ideally, strategies at least close to optimal can be found.

How to decide in case there is no single optimal strategy

Solution 1 (honest): *If we cannot decide, we cannot decide.*

- There are situations with too little information, or too much information that cannot be classified, to make well-founded decisions.
- Ideally, strategies at least close to optimal can be found.

Solution 2 (theoretical): *Consider a broader inference model.*

- A decision-theoretic approach (e. g. gains / losses following (1) – (2)) may be too narrow.
- More general approaches may not even lead to probabilities.

(D. Bernoulli 1738)

How to decide in case there is no single optimal strategy

Solution 1 (honest): *If we cannot decide, we cannot decide.*

- There are situations with too little information, or too much information that cannot be classified, to make well-founded decisions.
- Ideally, strategies at least close to optimal can be found.

Solution 2 (theoretical): *Consider a broader inference model.*

- A decision-theoretic approach (e. g. gains / losses following (1) – (2)) may be too narrow. (D. Bernoulli 1738)
- More general approaches may not even lead to probabilities.

Solution 3 (practical): *Restrict possible strategies*

- Classical: significance levels, invariance, minimax, ...
  Bayesian: Invent a prior, e. g. uniform distribution. (Laplace 1814)
- Simplify model by removing information
- This is somewhat arbitrary – why don't those decision criteria appear in the original model?

Recommendations

- If there is prior information that justifies a unique prior distribution, consider Bayesian methods.
- If there is prior information that cannot be captured in a single probability distribution, consider a range of reasonable prior distributions and check if the conclusions essentially remain the same.
- If there is no prior information at all (or one does not want to use such information) the results have a descriptive relevance, but general conclusions are hard to justify; even classical inferential procedures assume the considered alternatives are reasonably possible.

Summary

- Statistical decision theory justifies the separation of ignorance into probabilistic and completely unknown parts:
  - Classical: parameter unknown, sample probabilistic
  - Bayesian: everything probabilistic
  - in between: imprecise priors, robust Bayesian approaches, ...

- In many situations there is no unique optimal strategy. This leads to various competing solutions:
  - Classical solutions are easy to apply and to communicate in standard situations, but justifications are weak.
  - Bayesian solutions impose rather arbitrary precise priors.

  Recommendation: Consider a range of reasonable priors.

- Optimal strategies can exist even in sequential, imprecise settings.

Historical references

Fisher, 1921. *On the mathematical foundations of theoretical statistics.*

Kolmogorov, 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung.*

Wald, 1950. *Statistical decision functions.*

Ramsey, 1926. *Truth and probability.*

de Finetti, 1937. *La prévision: ses lois logiques, ses sources subjectives.*

Bernoulli (J.), 1713. *Ars conjectandi.*

Bernoulli (D.), 1738. *Specimen theoriae novae de mensura sortis.*

Neyman, Pearson, 1933. *On the problem of the most efficient tests of statistical hypotheses.*

Neyman, 1934. *Two different aspects of the representative method.*

Laplace, 1814. *Essai philosophique sur les probabilités.*