Probabilistic reasoning with uncertain evidence

Jiří Vomlel

LISP VŠE Praha

ÚTIA AV ČR

Probabilistic belief revision with new evidence

- We conducted a study on smoking S and lung cancer C.
- We get the counts $n(S = s, C = c)$ for $s, c \in \{yes, no\}$.
- Let $n = \sum_{s,c \in \{yes,no\}} n(S = s, C = c)$.
- We estimate probabilities $P(S = s, C = c) = \frac{n(S=s,C=c)}{n}$ for $s, c \in \{yes, no\}$.
- What is the probability of a person with lung cancer?
  $P(C = yes) = P(S = yes, C = yes) + P(S = no, C = yes)$
- We learn that the person we see is a smoker - we get an additional evidence $e = (S = yes)$.
- What is the probability of this person having lung cancer?
  $P(C = yes \mid e) = \frac{P(C=yes,e)}{P(e)} = \frac{n(S=yes,C=yes)}{n(S=yes)}$

The problem with uncertain evidence

- How should we revise our beliefs with uncertain evidence?

Examples

- “Based on what the physician sees on an x-ray she believes it is highly probable that the patient has cancer.”
- “The color of this cloth seems red, but I’m not sure, since it is quite dark and it is far away.”
- “The alarm in my house went on. There is a chance of burglary, but I know that reliability of the sensors is only 60%.”
- “An intelligence agency got report from an agent. They know that this source is usually very reliable.”

Our goal is to use the standard probability framework (with frequentist or subjective belief interpretation) to deal with uncertain evidence (in cases where it is suitable).

Reliability of information sources

A ... a variable of our interest,

T ... a test, a report, or an observation.

accuracy	$P(A = T) = \frac{tp+tn}{tp+tn+fp+fn}$
positive predictive value (precision)	$P(A = yes \mid T = yes) = \frac{tp}{tp+fp}$
negative predictive value	$P(A = no \mid T = no) = \frac{tn}{fn+tn}$
true positive rate (recall or sensitivity)	$P(T = yes \mid A = yes) = \frac{tp}{tp+fn}$
true negative rate (specificity)	$P(T = no \mid A = no) = \frac{tn}{fp+tn}$
false positive rate	$P(T = yes \mid A = no) = \frac{fp}{fp+tn}$
false negative rate	$P(T = no \mid A = yes) = \frac{fn}{tp+fn}$

Belief revision based on sources' reliability

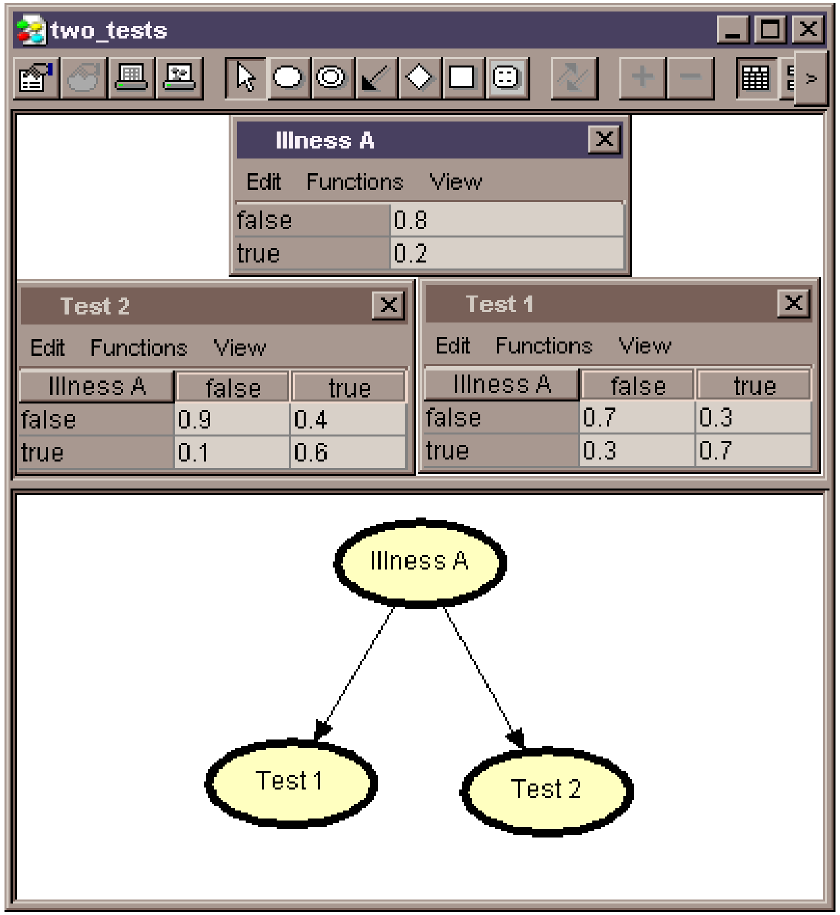
- sensitivity of T, i.e. $P(T = yes \mid A = yes) = P(t^+ \mid a^+)$
- specificity of T, i.e. $P(T = no \mid A = no) = P(t^- \mid a^-)$, and
- an observed result of T

Virtual evidence method (Pearl, 1988)

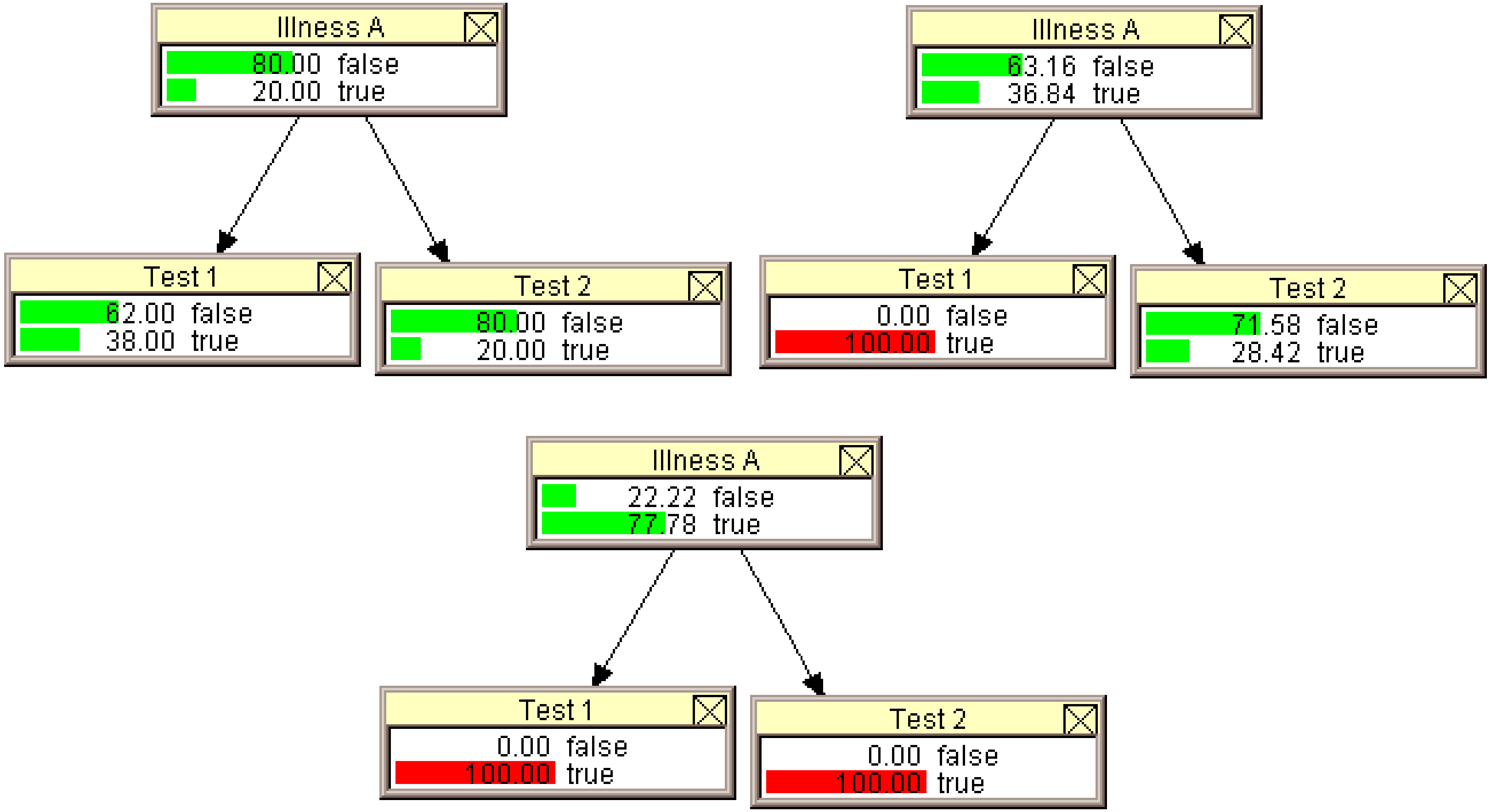
$$P(a^+ \mid t^+) = \frac{P(t^+ \mid a^+) \cdot P(a^+)}{P(t^+)}$$

$$= \frac{P(t^+ \mid a^+) \cdot P(a^+)}{P(t^+ \mid a^+) \cdot P(a^+) + (1 - P(t^- \mid a^-)) \cdot P(a^-)}$$

Example of two conditionally independent tests



Example of two conditionally independent tests



Probability kinematics (Jeffrey, 1988)

Assume

Ω	a finite set of elementary events
$\mathbf{A}$	set of all subsets of Ω
P, Q	two probability distributions on $\mathbf{A}$
$e_1, \ldots, e_n$	a set of mutually exclusive and exhaustive events

Definition

A probability distribution Q is said to come from P by probability kinematics on $e_1, \ldots, e_n$ if for every event $a \in \mathbf{A}$

$$P(a \mid e_i) = Q(a \mid e_i), \text{ for } i = 1, \ldots, n \ .$$

Jeffrey's rule

Assume

P	a probability distribution on $\mathbf{A}$
$e_1, \ldots, e_n$	a set of mutually exclusive and exhaustive events such that $P(e_i) > 0$ for $i = 1, \ldots, n$
$Q(e_i)$	posterior probability of event e_i, for $i = 1, \ldots, n$. Note that $\sum_i Q(e_i) = 1$.

Definition

A probability distribution Q comes from P by Jeffrey's rule if for all $a \in \mathbf{A}$

$$Q(a) \quad = \quad \sum_{i=1}^{n} Q(e_i) \cdot P(a \mid e_i) \ .$$

Jeffrey's rule obeys probability kinematics

Theorem

Probability distribution Q given by Jeffrey's rule is the only distribution:

- having $Q(e_i) = P(e_i)$ and
- obtained by probability kinematics from P on $e_1, \ldots, e_n$.

Observation variable (Valtorta, Kim, and Vomlel, 2002)

O	observation variable with states $o_1, \ldots, o_n$
o_i	state of O corresponding to event e_i, for $i = 1, \ldots, n$

$P(O = o_i \mid a) = 1$ iff $a = e_i$ for $i = 1, \ldots, n$

$Q(O = o_i) = Q(e_i)$

Assertion

A probability distribution Q comes from P by probability kinematics on $e_1, \ldots, e_n$ if for all $a \in \mathbf{A}$

$$Q(a) = \sum_{i=1}^{n} Q(O = o_i) \cdot P(a \mid O = o_i) \ .$$

Remark

This is what we get by standard probability update with new probability distribution $Q(O)$ on observation variable O.

Example: Tom and Mary go to a party (Pearl, 1990)

At the begining we know that

- there is a 50% chance that Tom goes to a party,
- there is a 50% chance that Mary goes to a party, and
- these two events are independent.

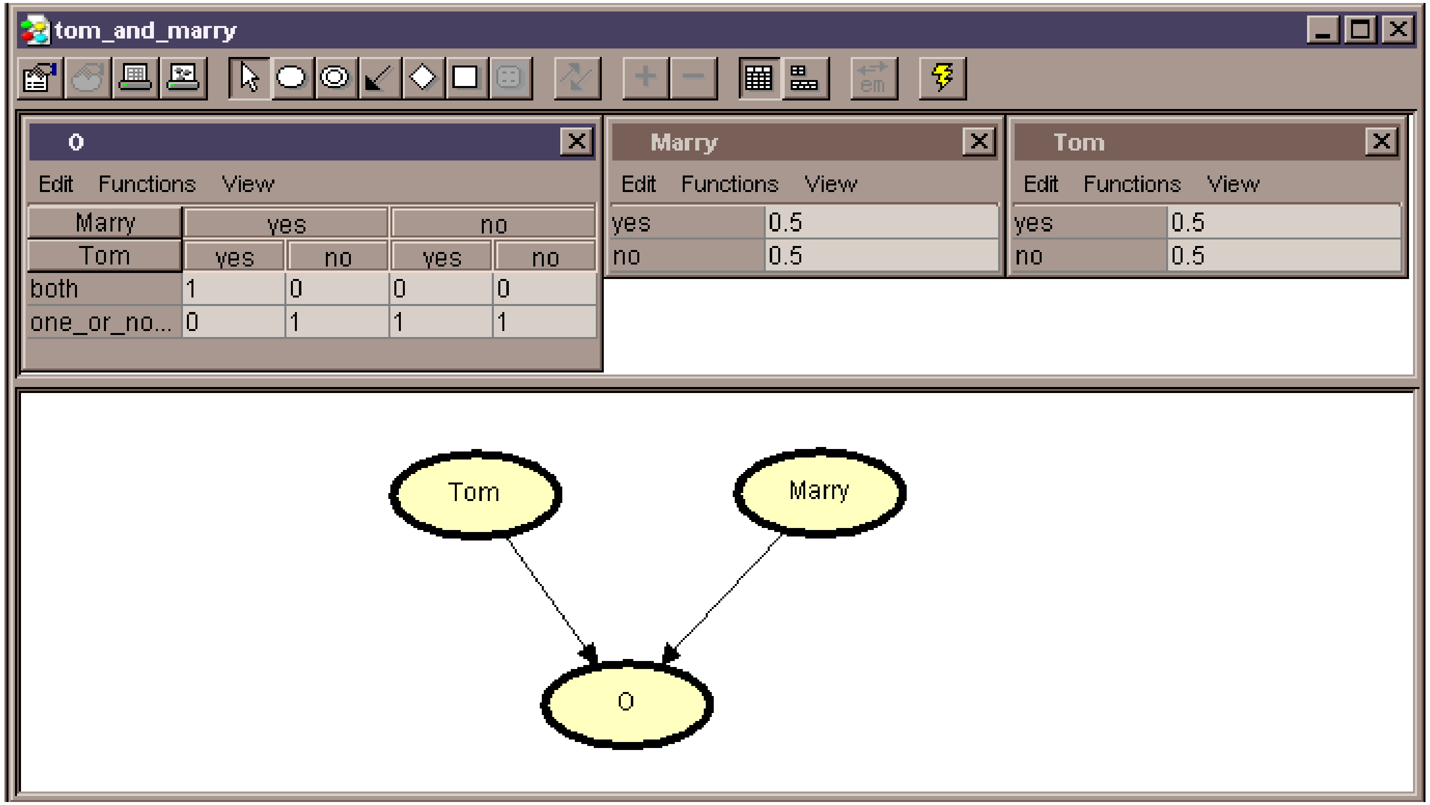
Latter we learn that

- *"Tom and Mary probably won't go together"* with probability 0.9.

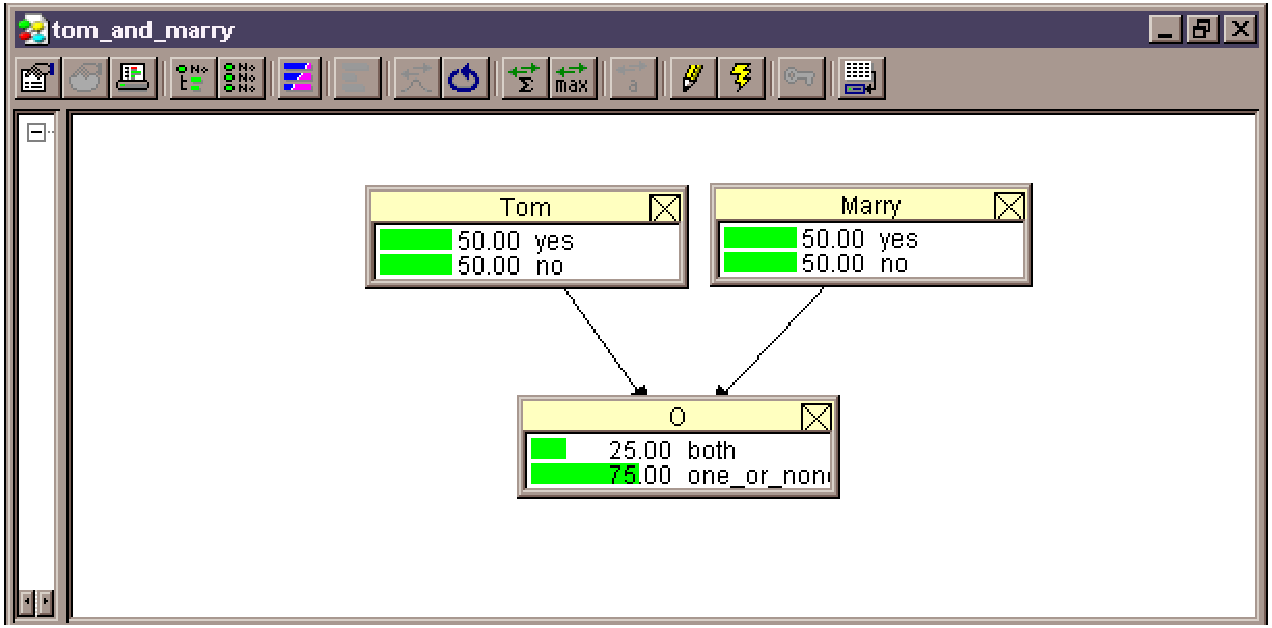
How should we revise probabilistic beliefs using the principle of probability kinematics?

Tom and Mary go to a party

model with an observational variable



Tom and Mary go to a party - initial beliefs

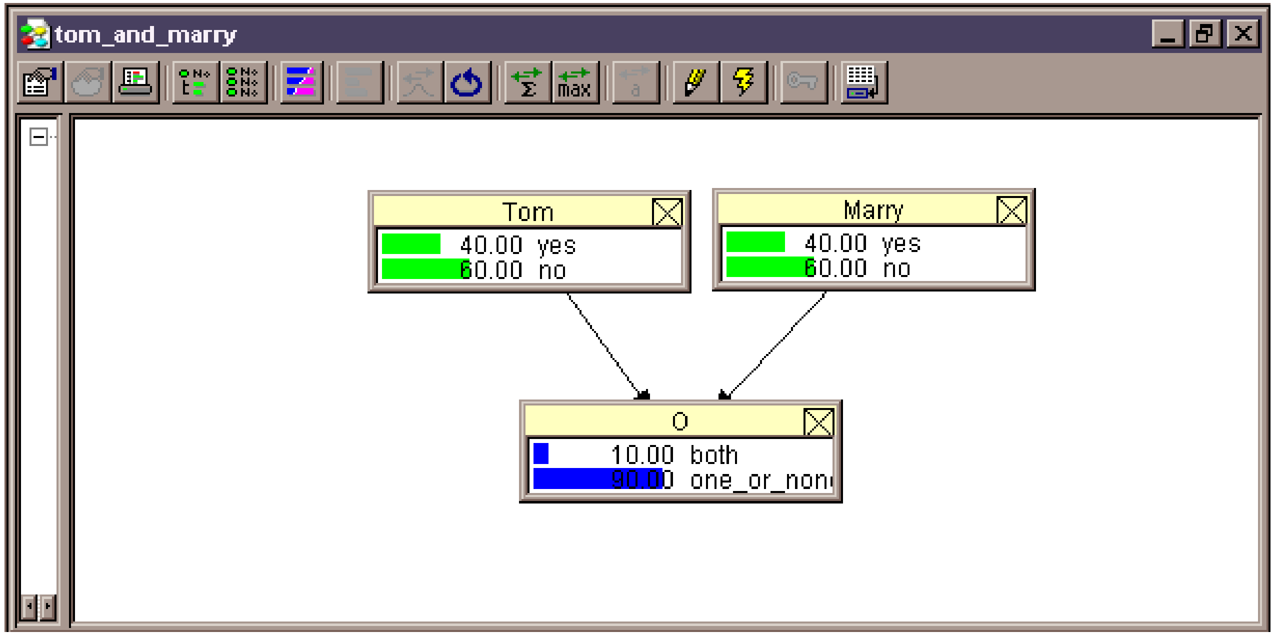


Tom and Mary go to a party - revised beliefs

$$Q(T = t, M = m)$$

$$= \sum_{i=1}^{2} Q(O = o_i) \cdot P(T = t, M = m \mid O = o_i)$$

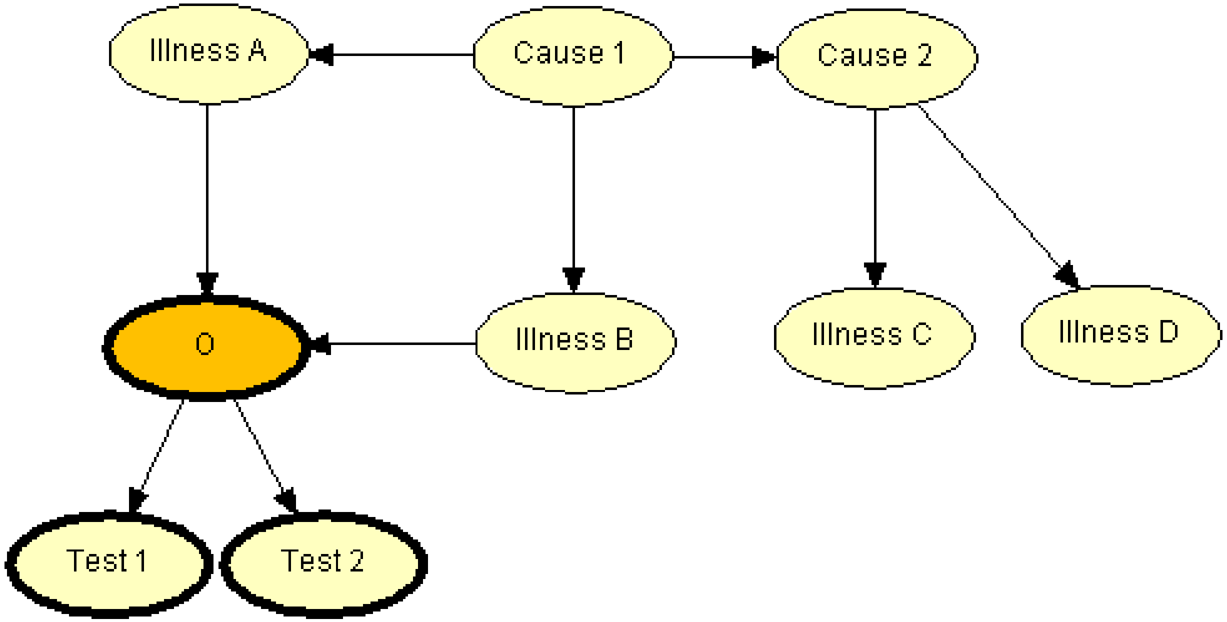
$$= \sum_{i=1}^{2} Q(O = o_i) \cdot \frac{1}{P(O = o_i)} \cdot P(O = o_i \mid T = t, M = m) \cdot P(T = t, M = m)$$



$Q(T, M)$	$T = y$	$T = n$
$M = y$	0.1	0.3
$M = n$	0.3	0.3

I-projection of P to the set of distributions with $Q(T = y, M = y) = 0.1$

Belief revision in a complex probabilistic model $P(V)$



Expert calculations:

$$Q(O) = P(O \mid T_1, \ldots, T_n)$$
$$= c \cdot P(O) \cdot P(T_1 \mid O) \cdot P(T_2 \mid O)$$

Belief revision in the model:

$$Q(V) = \sum_O P(V \mid O) \cdot Q(O)$$

Belief vs. Model revision

a fundamental difference

- The posterior probabilities after *belief revision* inform about the properties of a tested individual, a performed event, etc.
- The posterior probabilities after a *model revision* still correspond to the properties of the tested population of individuals, events, etc.

Model revision

revision of a model parameter

- for each parameter only one value
  $\longrightarrow$ *maximum likelihood estimation*
- posterior probabilities defined over the model parameters
  $\longrightarrow$ *Bayesian statistics*

Example - maximum likelihood estimate of a parameter

Two variables (their states are known in the tested population):

- test result T (positive/negative)
- illness A (sick/non-sick).

Likelihood of observed data D for a given model $P(T, A)$

$$L(P \mid D) \quad = \quad \prod_{(t,a)} P(T = t, A = a)^{n(t,a)}$$

Sensitivity r of test T was defined as $P(T = yes \mid A = yes)$.
Maximum likelihood estimate:

$$\hat{r} \quad = \quad \arg\max_r L(P \mid D)$$

Example - combining two maximum likelihood estimates

Two experts evaluated sensitivity r of a test. They computed their maximum likelihood estimate

$$\hat{r}_i = \frac{n_i(T = yes, A = yes)}{n_i(A = yes)} \text{ for } i = 1, 2.$$

$$L(P \mid D) = \prod_{(t,a)} P(T = t, A = a)^{n_1(t,a)} \cdot P(T = t, A = a)^{n_2(t,a)}$$

Maximum likelihood estimate of r is weighted average of $\hat{r}_1$ and $\hat{r}_2$

$$\begin{aligned} \hat{r} &= \frac{n_1(A = yes)}{n_1(A = yes) + n_2(A = yes)} \cdot \hat{r}_1 + \frac{n_2(A = yes)}{n_1(A = yes) + n_2(A = yes)} \cdot \hat{r}_2 \\ &= w_1(yes) \cdot \hat{r}_1 + (1 - w_1(yes)) \cdot \hat{r}_2 \end{aligned}$$

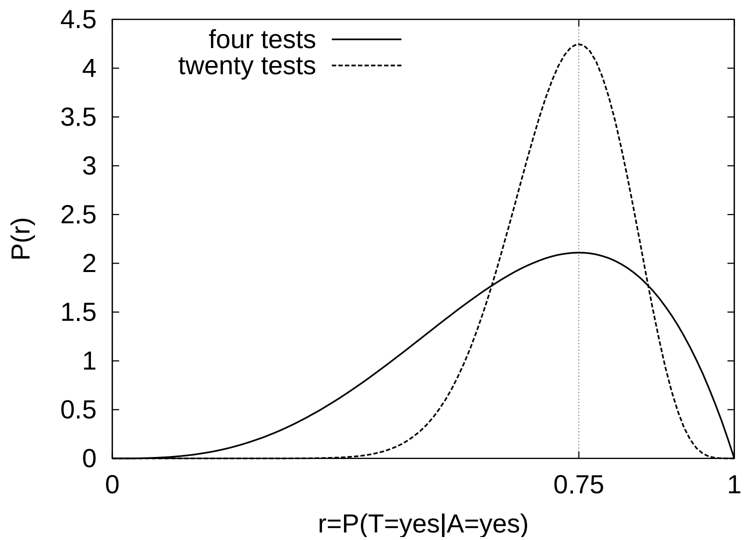
Example - posterior probability of model parameters

Instead of a single value $\hat{r}$ we take the whole distribution.

(1) $n_1(A = yes) = n_2(A = yes) = 2$

(2) $n_1(A = yes) = n_2(A = yes) = 10$

In both cases $\hat{r}_1 = 0.7$ and $\hat{r}_2 = 0.8$ and the prior distribution is uniform.



Summary

- Bayesian networks are suitable for reasoning with uncertainty.
- Standard efficient methods for belief revision with certain evidence are available (since 90’s).
- Different methods for belief revision with uncertain evidence can be used depending on what is the reported belief referring to.
- When beliefs are reported in natural language it may be tricky to find out what the number actually means.
- There is a fundamental difference between belief and model revision.
- Sometimes most likely values of model parameters are sufficient while in some situation it is more appropriate to use whole posterior probability distribution of parameter values.