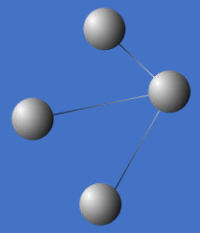
Making equivalent sample size an integral part of Bayesian network modelling practices

Steven Mascaro

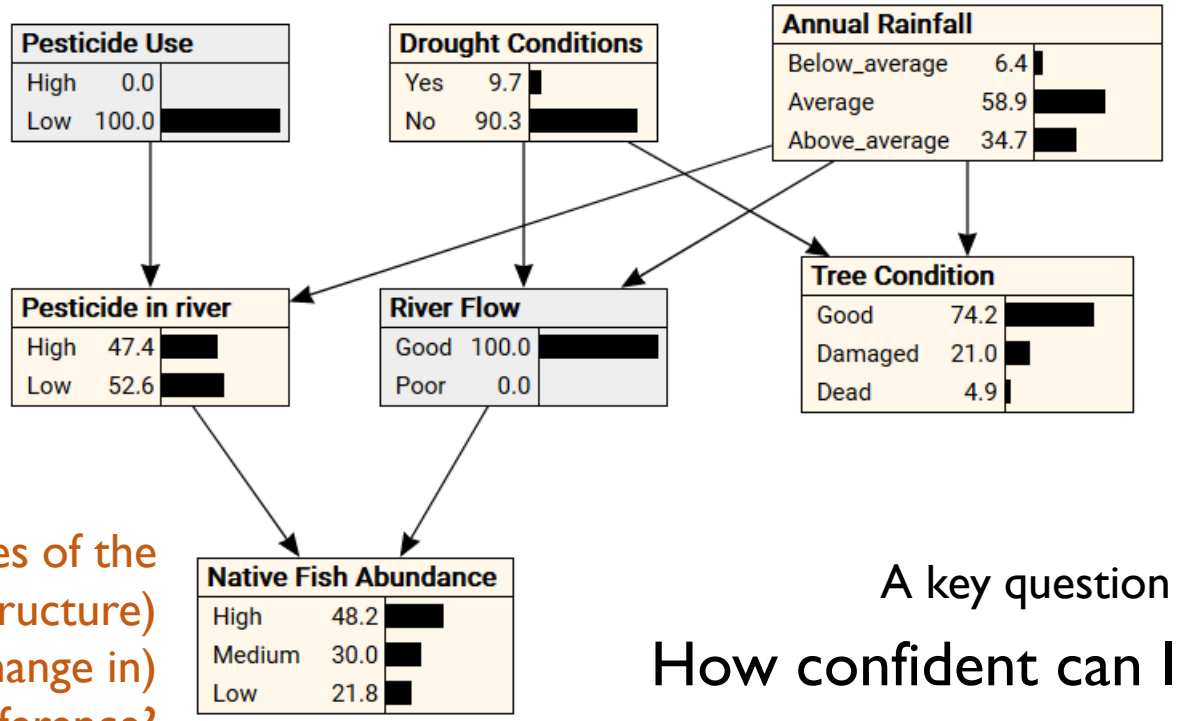


BAYESIAN
INTELLIGENCE

Confidence

BNs are very good at uncertainty…

…they're less helpful with the weight of evidence for our uncertainty



What support does this distribution have?

How robust is this inferred distribution?

To what degree are the BN's beliefs here due to data vs expert estimates?

What features of the parameters (& structure) led to this (change in) inference?

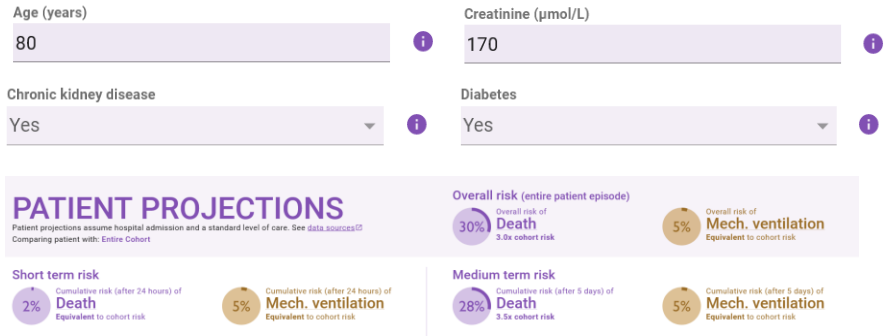
A key question that often arises is:

How confident can I be in the BN's outputs?

What do we mean by confidence?

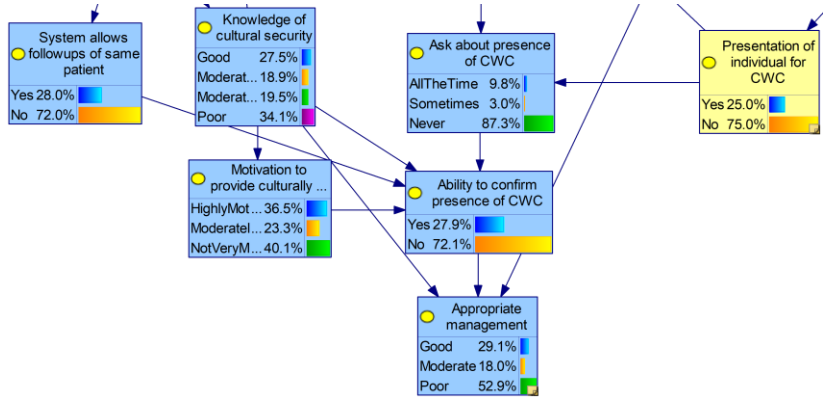
Often: How likely is the BN's posterior distribution to match reality?

(Past, present or future)



COVID-I

(Health calculator based on observational data)



APPLE-BN

(Implementation support informed by experts/surveys)

We will:

- Suggest **Equivalent Sample Size (ESS)** as a <u>partial</u> answer
- Provide a method for **estimating conditional ESS**
- Explore its **correctness**, **intuitiveness** and whether it really helps with **confidence**

BAYESIAN INTELLIGENCE

Equivalent Sample Size (ESS)

We will use this definition:

> The estimated weight of evidence for a probability distribution, measured in number of samples.

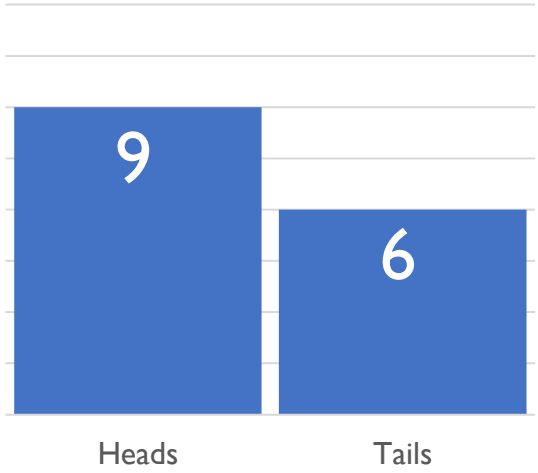
Measured in "number of samples" – doesn't need to be a real sample

It's an estimate, and so clearly subject to uncertainty

It seems too simple to cover all use cases (like experts), but maybe not

ESS: simple example

We flip a coin 15 times and get:



With no priors, we estimate:

P(Heads)=0.6

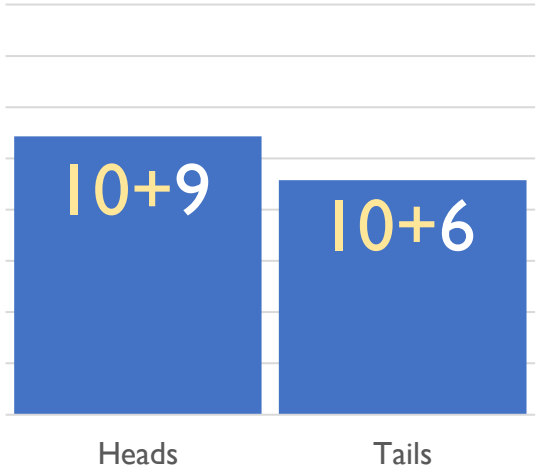
P(Tails)=0.4

By our definition, the ESS of this Bernoulli is 15

Bernoullis don't specify an ESS – we're just noting *this* case used an ESS of 15

ESS: simple example

We use a (Bayesian) prior count of 10 for both heads and tails:



P(Heads)=0.543

P(Tails)=0.457

15 new samples + 20 samples prior

By our definition, we estimate the ESS of this Bernoulli to be 35

Formal ESS

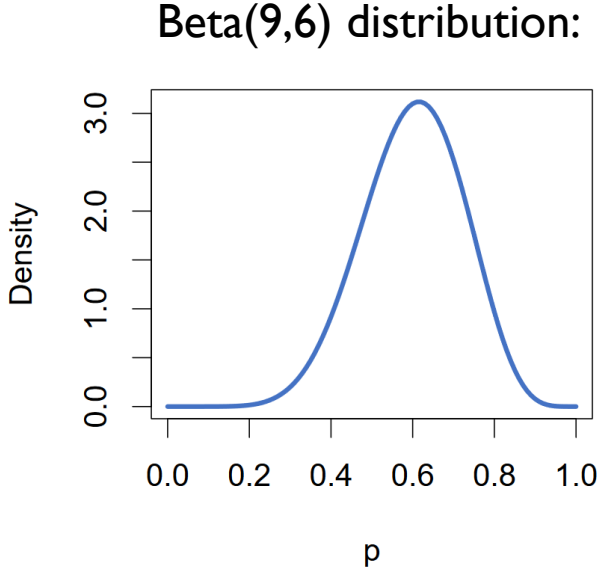
A Beta distribution can be interpreted as the *probability distribution over a Bernoulli*, once we assume an ESS

We can even specify the Beta directly with the Head/Tail counts:

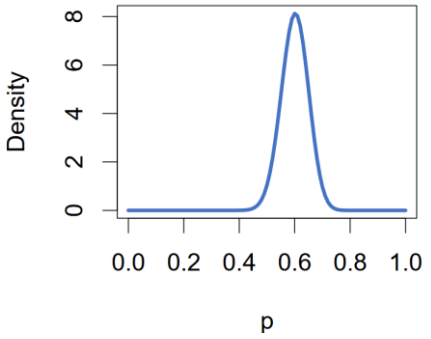
Beta(9,6)

A Dirichlet is just a Beta with more states. e.g.:

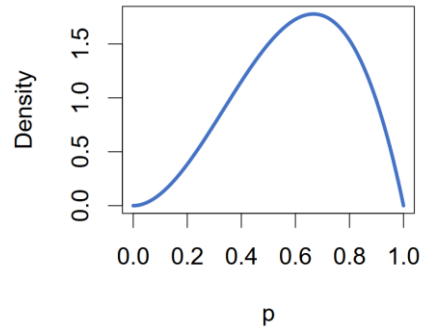
Dirichlet(9,1,5,6) for 4 states



ESS: 15



ESS: 90

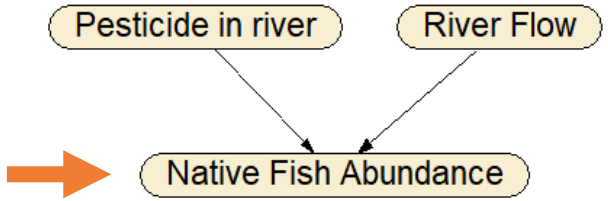


ESS: 5

The only difficulty is the visualisation is multi-dimensional ☹

CPTs and learning

Parameter learning refresher:



Plain CPT

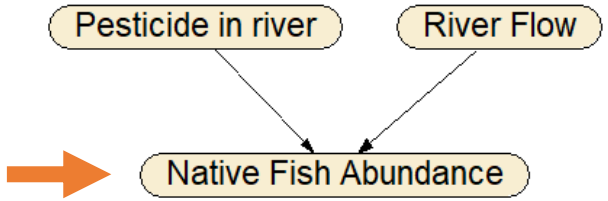
Pesticide i...	River Flow	High	Medium	Low
High	Good	18.421	39.474	42.105
High	Poor	[illegible]	[illegible]	[illegible]
Low	Good	75	21.429	3.571
Low	Poor	20	12	68

Dirichlets (with ESS)

Pestici...	River ...	High	Medium	Low	ESS
High	Good	7	15	16	38
High	Poor	1	2	[illegible]	21
Low	Good	21	6	1	28
Low	Poor	5	3	17	25

CPTs and Learning

Parameter learning refresher:



Plain CPT

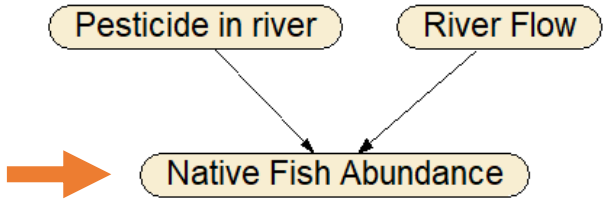
Pesticide i...	River Flow	High	Medium	Low
High	Good	18.421	39.474	42.105
High	Poor	[illegible]	[illegible]	[illegible]
Low	Good	75	21.429	3.571
Low	Poor	20	12	68

Dirichlets (with ESS)

Pestici...	River ...	High	Medium	Low	ESS
High	Good	7	15	17	39
High	Poor	1	2	[illegible]	21
Low	Good	21	6	1	28
Low	Poor	5	3	17	25

CPTs and learning

Parameter learning refresher:



Plain CPT

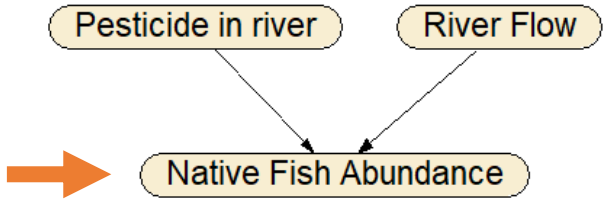
Pesticide i...	River Flow	High	Medium	Low
High	Good	17.949	38.462	43.59
High	Poor	[illegible]	[illegible]	[illegible]
Low	Good	[illegible]	[illegible]	[illegible]
Low	Poor	20	12	68

Dirichlets (with ESS)

Pestici...	River ...	High	Medium	Low	ESS
High	Good	7	15	17	39
High	Poor	1	2	[illegible]	21
Low	Good	[illegible]	6	1	28
Low	Poor	5	3	17	25

CPTs and learning

Parameter learning refresher:



Plain CPT

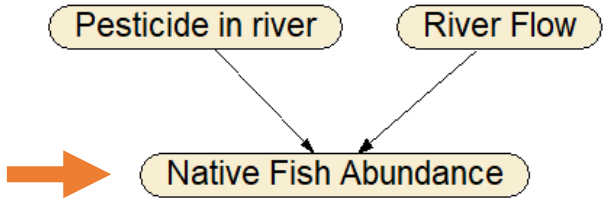
Pesticide i...	River Flow	High	Medium	Low
High	Good	17.949	38.462	43.59
High	Poor	4.762	9.524	85.714
Low	Good	[illegible]	[illegible]	[illegible]
Low	Poor	20	12	68

Dirichlets (with ESS)

Pestici...	River ...	High	Medium	Low	ESS
High	Good	7	15	17	39
High	Poor	1	2	18	21
Low	Good	[illegible]	6	1	28
Low	Poor	6	3	17	26

CPTs and learning

Parameter learning refresher:



Plain CPT

Pesticide i...	River Flow	High	Medium	Low
High	Good	17.949	38.462	43.59
High	Poor	4.762	9.524	85.714
Low	Good	[illegible]	[illegible]	[illegible]
Low	Poor	23.077	11.538	65.385

Dirichlets (with ESS)

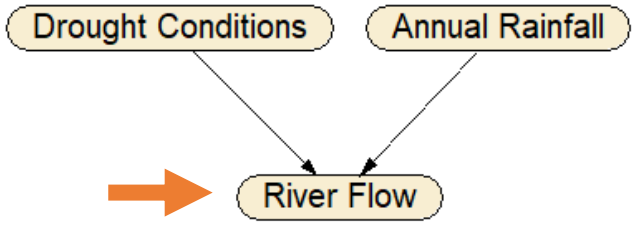
Pestici...	River ...	High	Medium	Low	ESS
High	Good	7	15	17	39
High	Poor	1	2	18	21
Low	Good	[illegible]	6	1	28
Low	Poor	6	3	17	26

Using ESS after training

After training, we tend to ignore ESS…

…but it's essential to interpreting the outputs

Example:

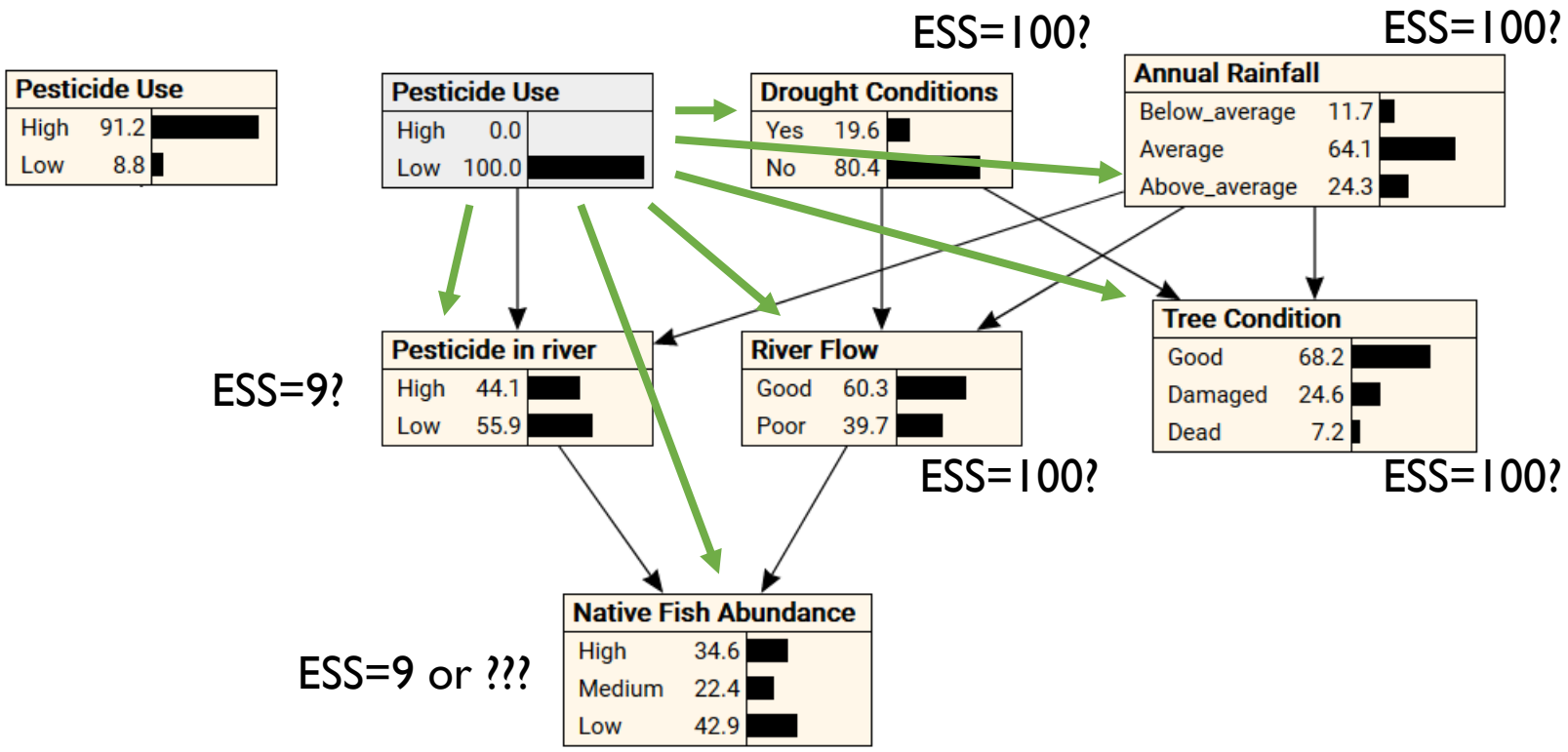


Drought Conditions	Annual Rainfall	Good	Poor	ESS
Yes	Below average	33.333	66.667	3
Yes	Average	21.429	78.571	14
Yes	Above average	50	50	8
No	Below average	33.333	66.667	12
No	Average	63.636	36.364	55
No	Above average	95	5	20

Conditional ESS

What might the ESS be for *any* node, given we put in some evidence *anywhere else*?

Let's look at a model trained on 100 cases:



Goals

- Estimate the conditional ESS for every node
- Good for expert models, not just trained models
- Good for latent variables
- Show how much ESS is due to priors, data, experts, etc.

Non-goal:

- Good for incorrect structures
  (structure assumed true)
  (actually, just need statistical equivalence)

Initial alternatives considered

Joint probability of data
×
number of cases

Too low
(d-separated nodes still affect estimate)

Joint probability of d-connected evidence
×
number of cases

Still too low
(same effect no matter how long the chain)

MI% weighted probability of evidence
×
number of cases

Much too high
(MI% falls off too quickly)

Method

- Assume all CPT rows in the BN have an ESS
  (typically the case in a trained BN – in Netica, at least)
- Assume evidence has been entered
- For *n* iterations:
  - For every row in every node:
    - Sample the Dirichlet
    - Replace the row with the sample
  - Do a belief update and store the beliefs for all nodes
- Compute the variance of the beliefs for each node
- Estimate the node's ESS from its belief variance

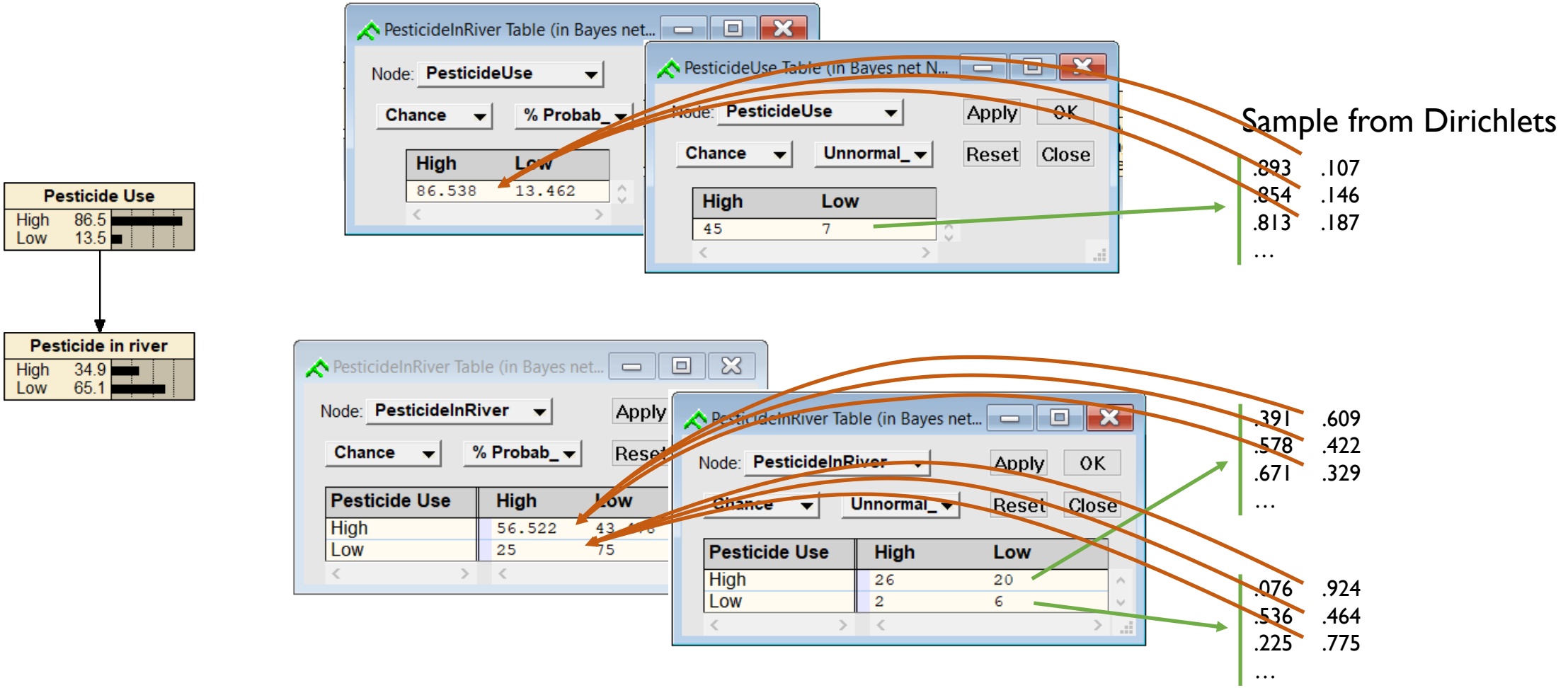
Sampling a Dirichlet produces a probability vector

e.g. 4 samples of a Dirichlet <5,1,9> (ESS=15):
.39 .08 .53
.43 .05 .52
.23 .09 .68
.12 .02 .86

4 samples of a Dirichlet <50,10,90> (ESS=150):
.28 .05 .67
.33 .04 .62
.31 .06 .64
.34 .06 .60

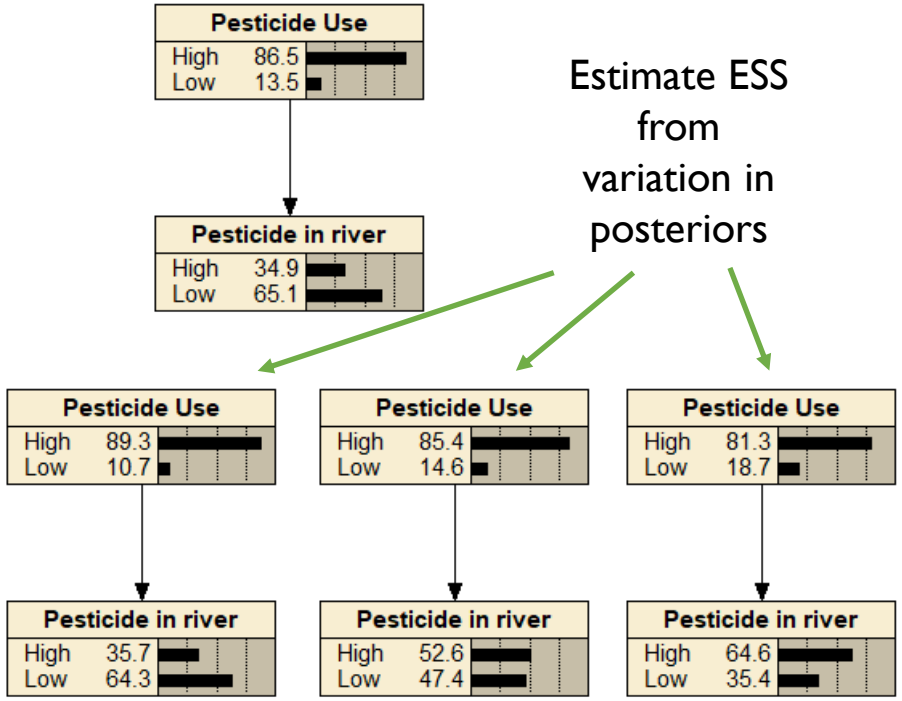
Dirichlet sample variance and ESS are inversely related!

Illustration of method

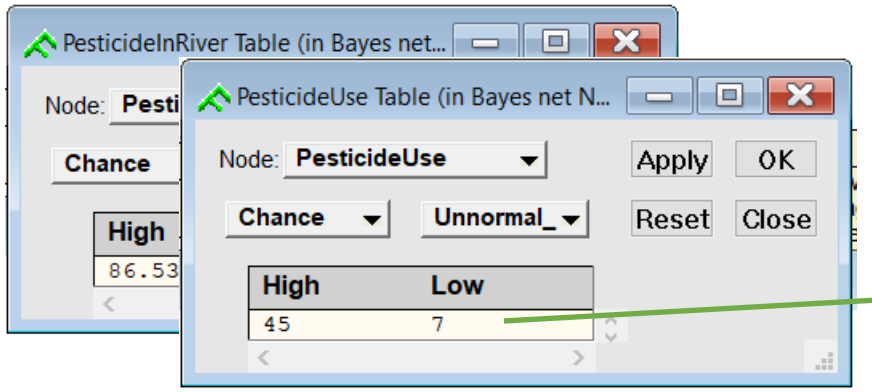


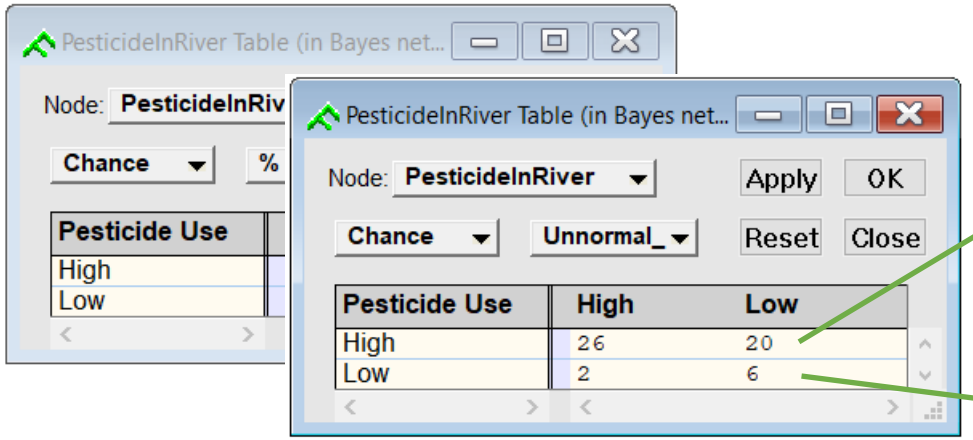
BAYESIAN INTELLIGENCE

Illustration of method



Estimate ESS from variation in posteriors





Sample from Dirichlets

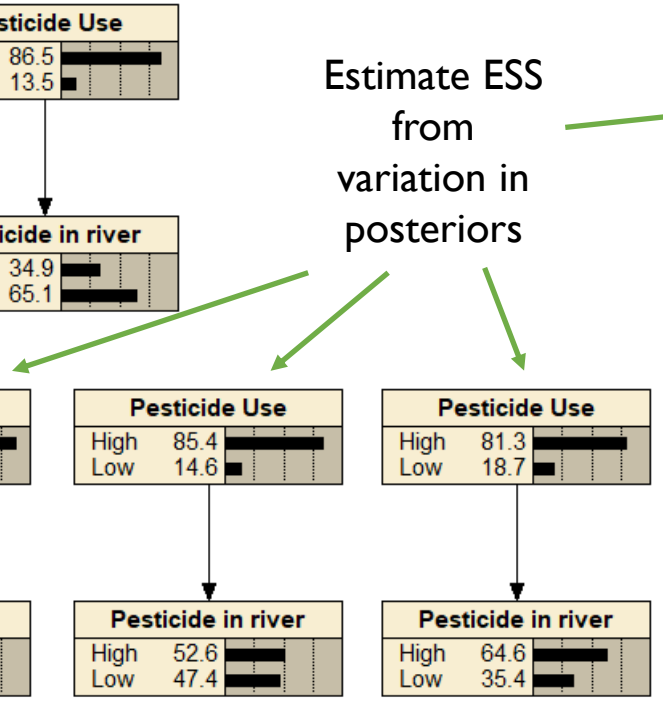
.893	.107
.854	.146
.813	.187
…	

.391	.609
.578	.422
.671	.329
…	

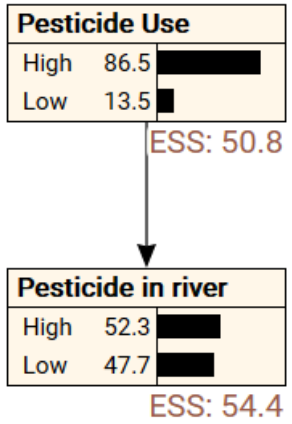
.076	.924
.536	.464
.225	.775
…	

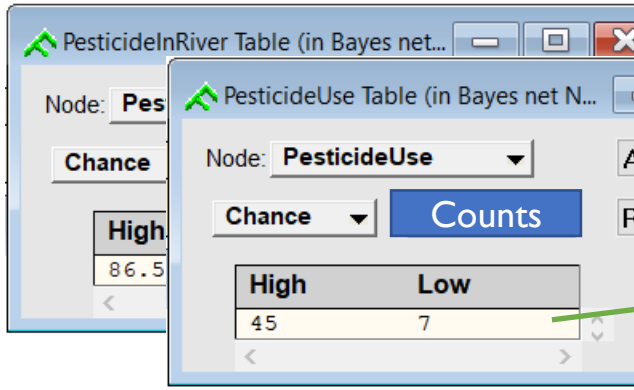
BAYESIAN INTELLIGENCE

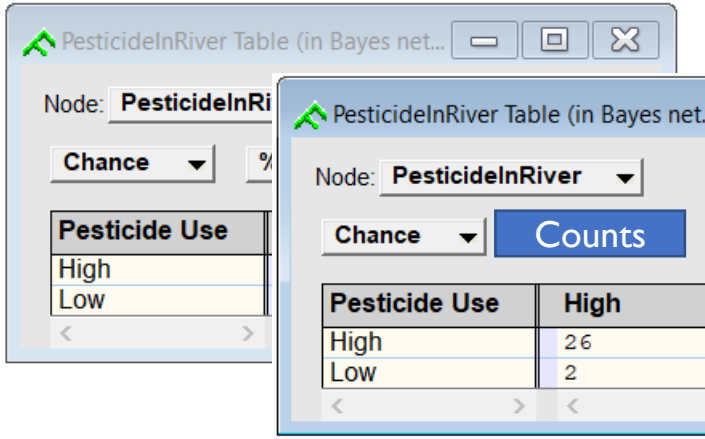
Illustration of method



Estimate ESS from variation in posteriors



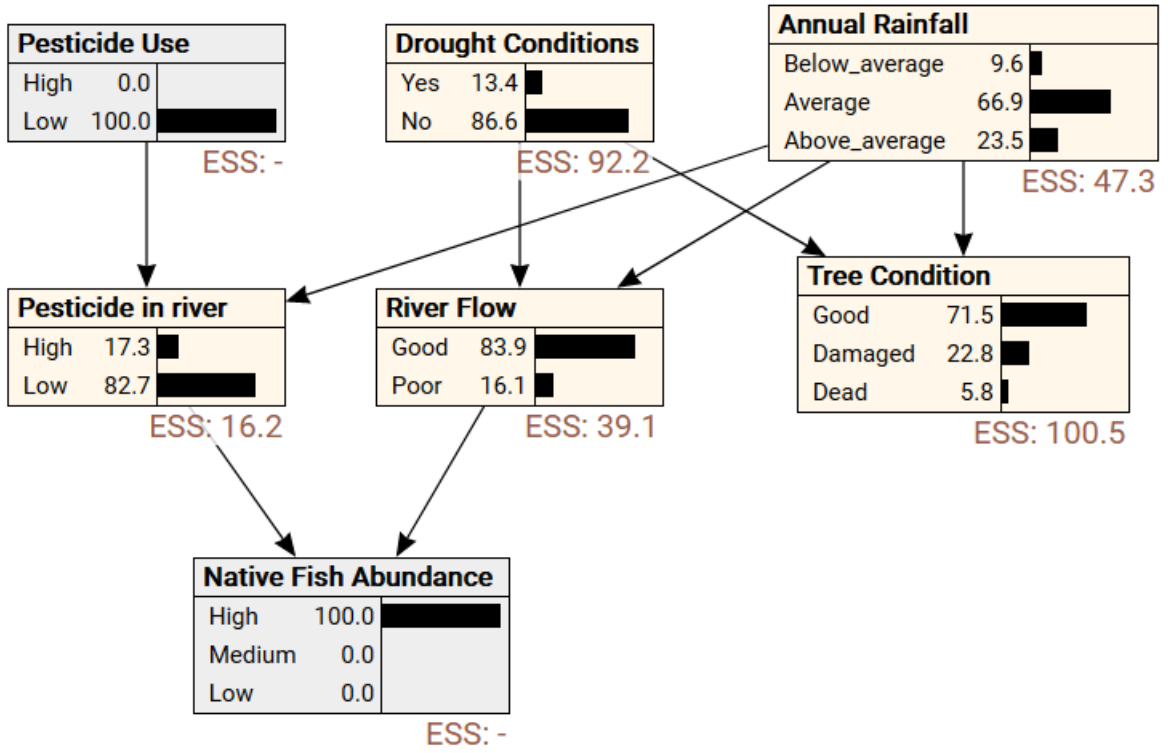




BAYESIAN INTELLIGENCE

Example on larger (but still toy) BN

Demo



(Trained on 100 cases)

Does it match expectations?

In particular:

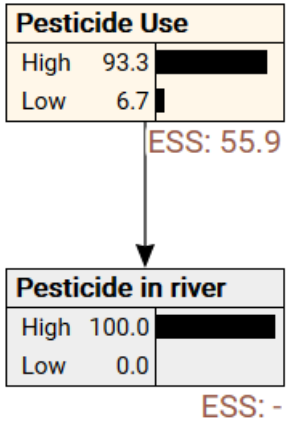
- Does it seem correct?
- Is it intuitive?
- Does it help us see if the inference will match the true distribution?

  (assuming the structure is correct!)

Does it match expectations?

- Does it seem correct?
- Is it intuitive?
- Does it help us see if the inference will match the true distribution?

2 node BN



High	Low
93	9

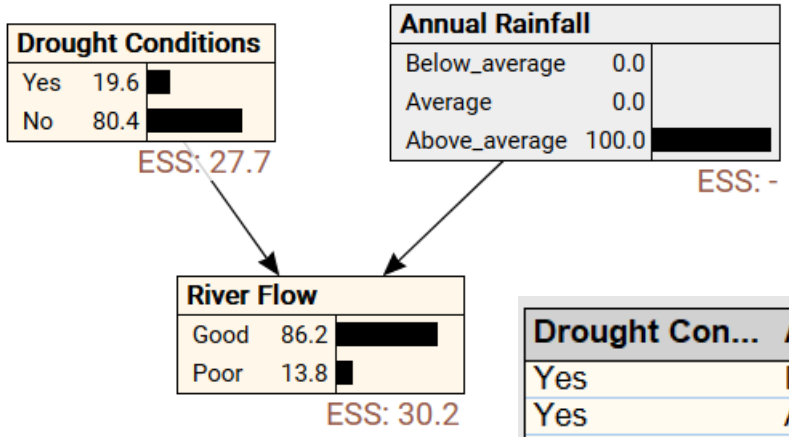
102

Pesticide Use	High	Low	
High	51	43	94
Low	4	6	10
	55	49	104

Does it match expectations?

- Does it seem correct?
- Is it intuitive?
- Does it help us see if the inference will match the true distribution?

2 parents



Drought Con...	Annual Rainfall	Good	Poor
Yes	Below average	1	2
Yes	Average	3	11
Yes	Above average	4	4
No	Below average	4	8
No	Average	35	20
No	Above average	19	1

ESS

3 15
14 29
8 28
12
55 87
20

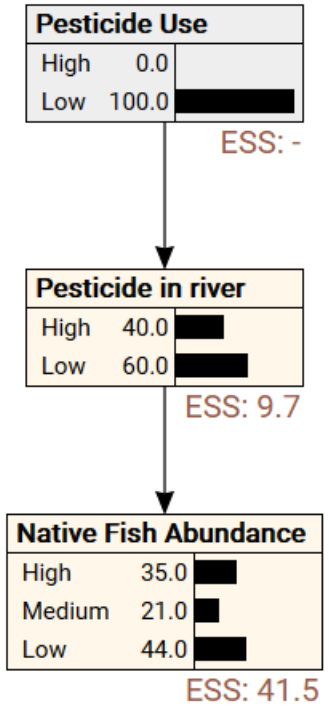
112

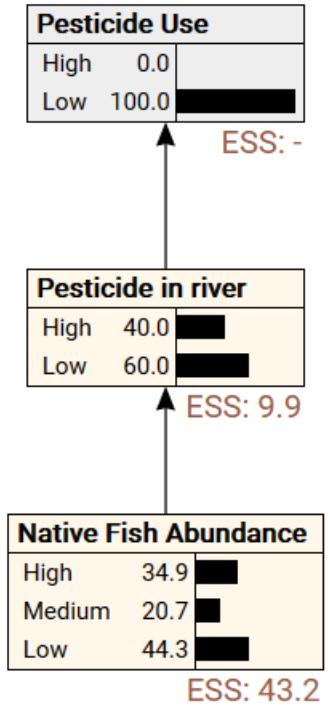
BAYESIAN INTELLIGENCE

Does it match expectations?

- Does it seem correct?
- Is it intuitive?
- Does it help us see if the inference will match the true distribution?

Chains

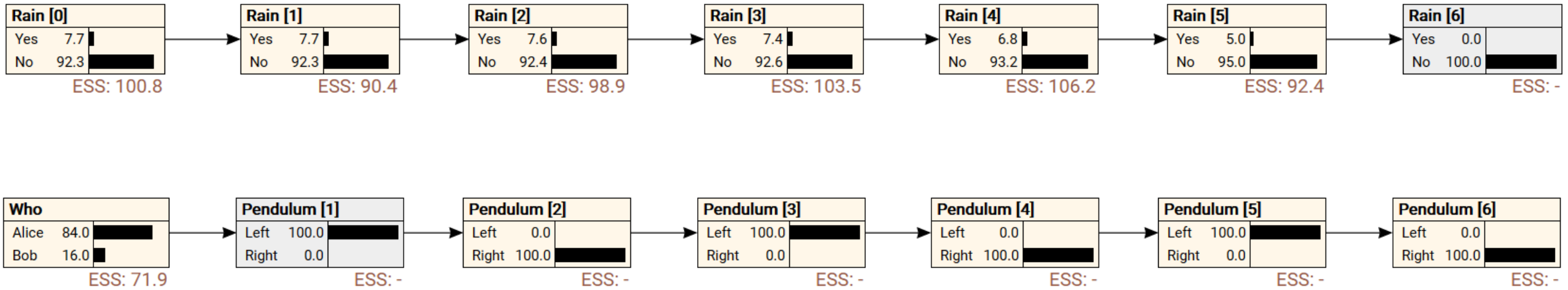




Does it match expectations?

- Does it seem correct?
- Is it intuitive?
- Does it help us see if the inference will match the true distribution?

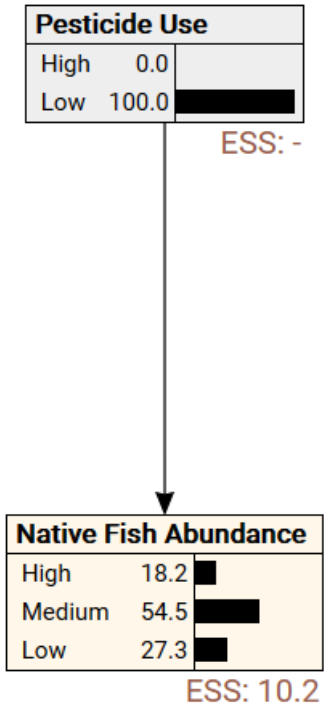
Long chains

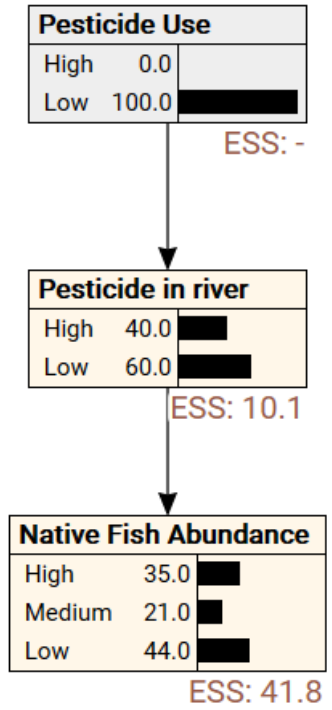


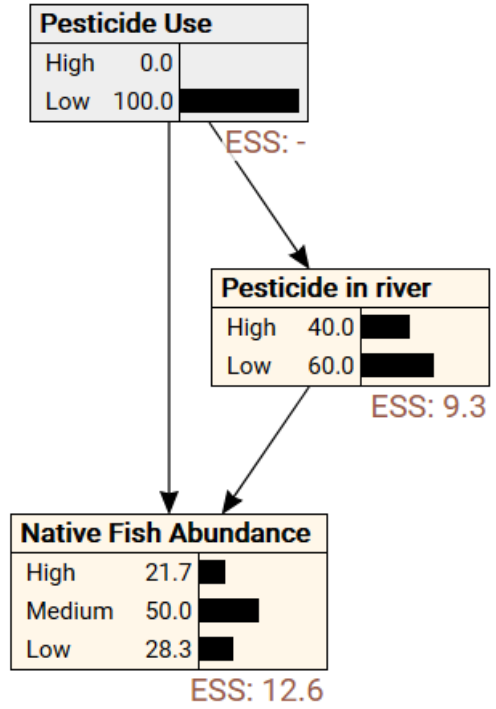
Does it match expectations?

- Does it seem correct?
- Is it intuitive?
- Does it help us see if the inference will match the true distribution?

3 node variations





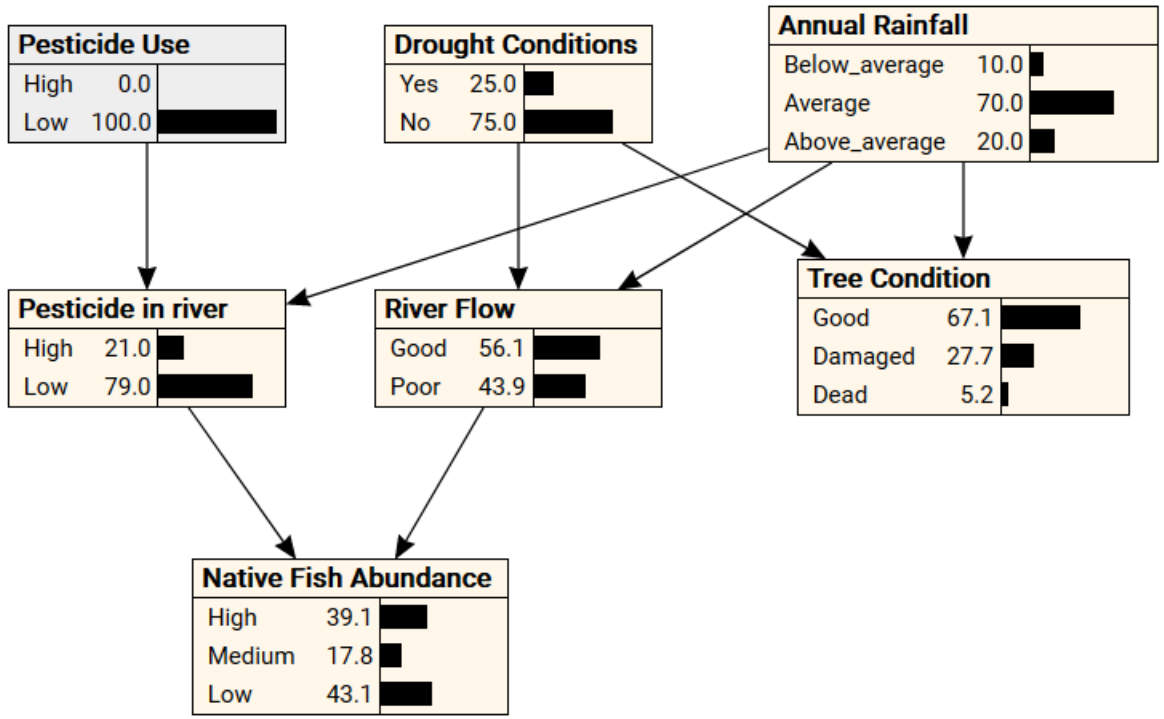


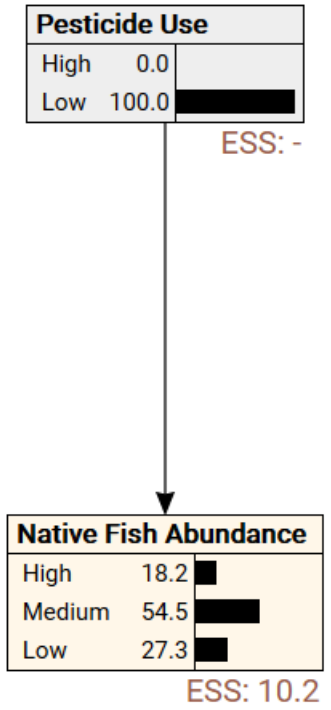
BAYESIAN INTELLIGENCE

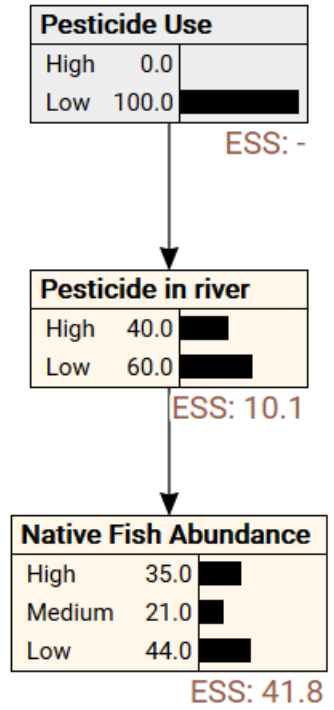
Does it match expectations?

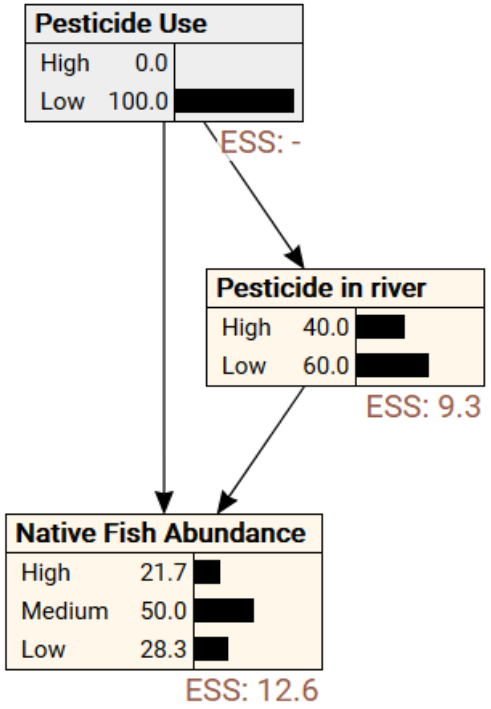
- Does it seem correct?
- Is it intuitive?
- Does it help us see if the inference will match the true distribution?

3 node variations





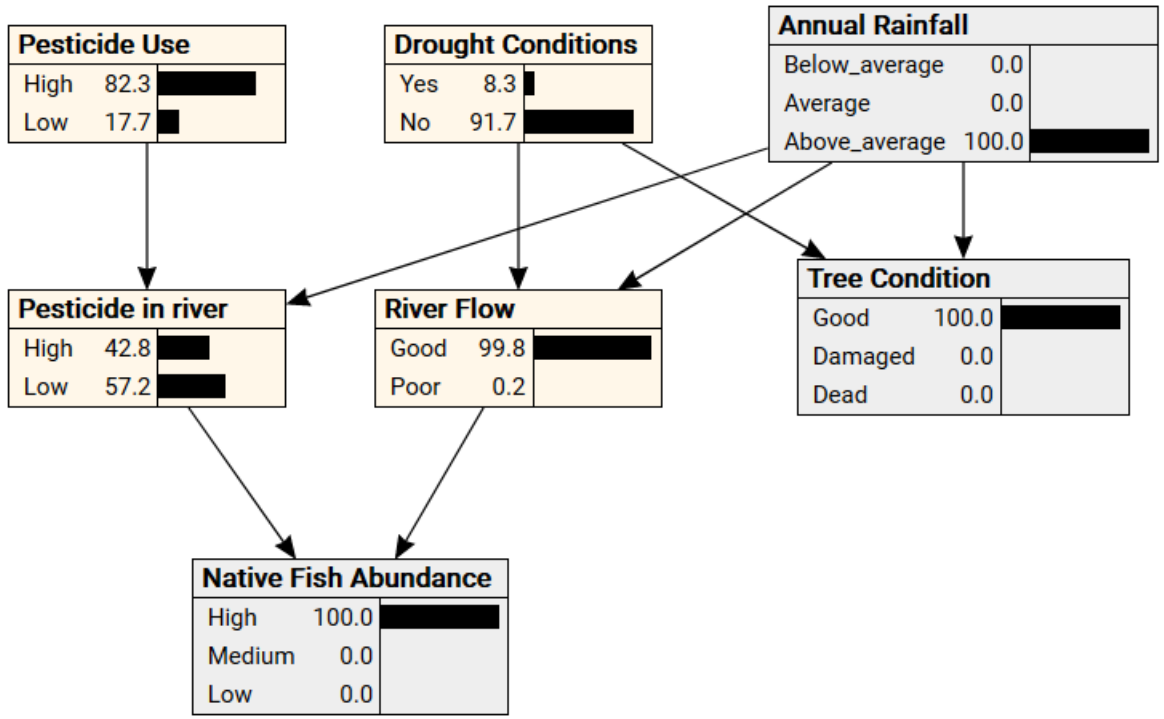




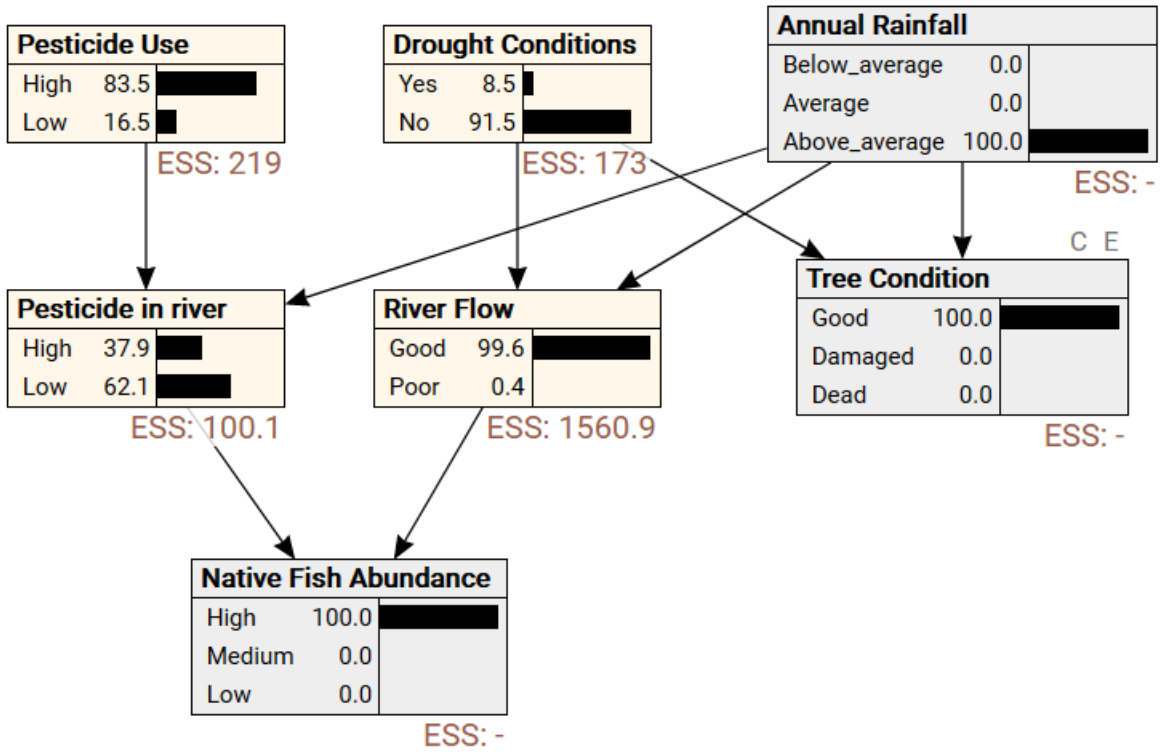
BAYESIAN INTELLIGENCE

Does it match expectations?

- Does it seem correct?
- Is it intuitive?
- Does it help us see if the inference will match the true distribution?



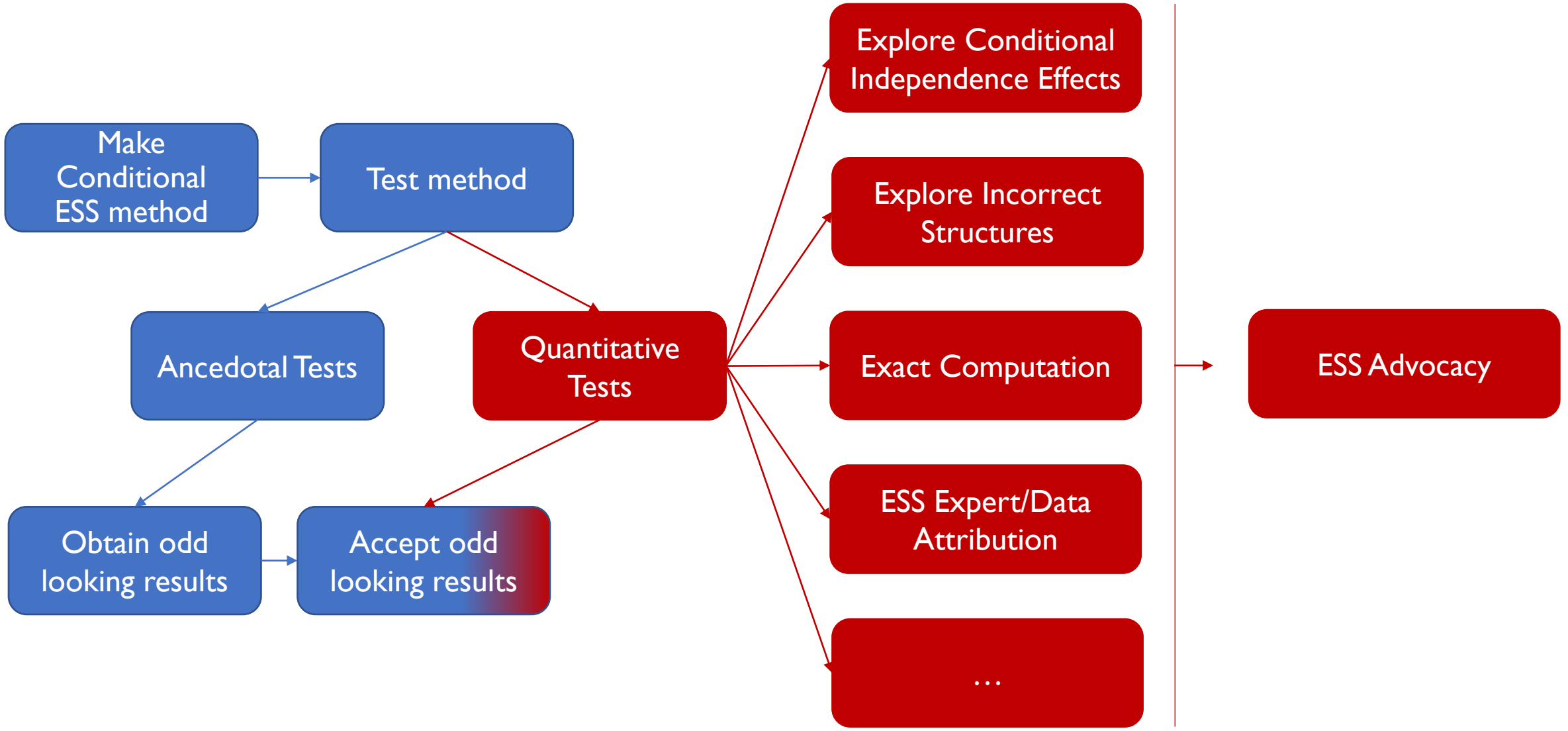
To the sample and beyond…



(Trained on 1000 cases)

BAYESIAN INTELLIGENCE

Current & Future Work



BAYESIAN INTELLIGENCE