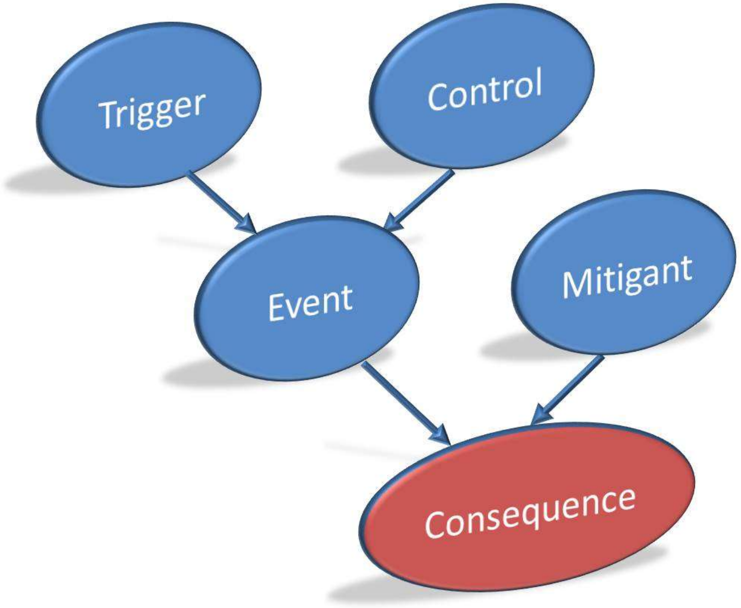
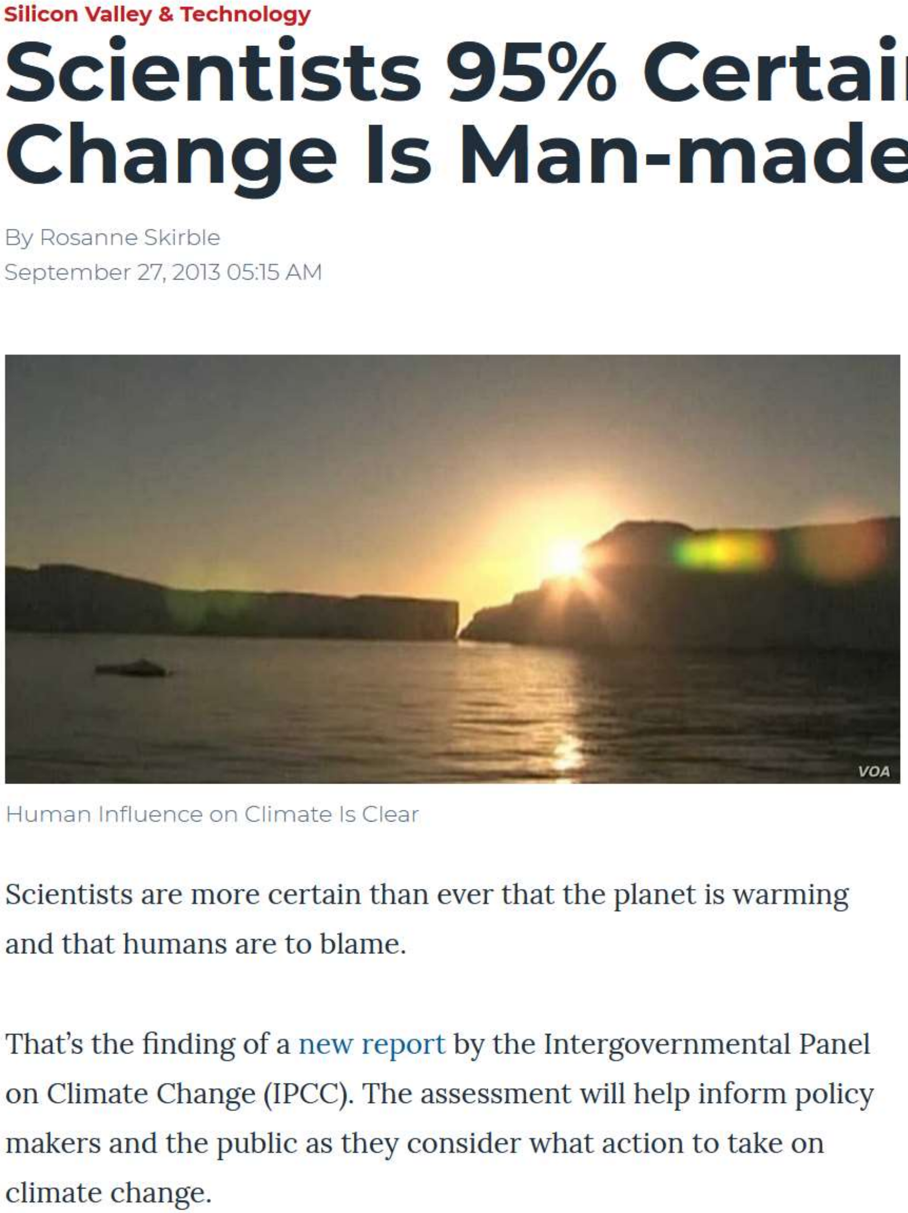


Risk and Decision Making for Data Science and AI

Lesson 4b Hypothesis testing: from classical to Bayesian

Norman Fenton

@ProfNFenton



Silicon Valley & Technology

Scientists 95% Certain Climate Change Is Man-made

By Rosanne Skirble
September 27, 2013 05:15 AM

Human Influence on Climate Is Clear

Scientists are more certain than ever that the planet is warming and that humans are to blame.

That's the finding of a new report by the Intergovernmental Panel on Climate Change (IPCC). The assessment will help inform policy makers and the public as they consider what action to take on climate change.

Does this mean:

- There is a 95% probability that most** recent global warming is man made
- 95% of scientists agree that most recent global warming is man made
- The probability of observing the recent global climate data if it were mostly man made is 95%
- The probability of observing the recent global climate data if it were not mostly man made is 5%
- Something else

** 'most', and 'mostly' here means "*at least half*"

See: https://probabilityandlaw.blogspot.com/2015/02/the-statistics-of-climate-change.html

Hypothesis testing

New drug 95% certain to cure Hepatitis C

Does this mean:

- 95% of Hepatitis C patients tested were cured with the drug
- There is a 95% probability that a patient with Hepatitis C will be cured if they take the drug
- 95% of doctors agree that the drug cures Hepatitis C
- There is a 95% probability of observing the success rate of the drug on patients tested so far if the drug is a cure for Hepatitis C
- There is a 5% probability of observing the success rate of the drug on patients tested so far if the drug is not a cure for Hepatitis C
- Something else

Classical hypothesis testing and *p*-values

We have a hypothesis ***H*** that we believe/hope might be true

It is often difficult/impossible to ‘prove’ a hypothesis is true

Instead, we get observed data ***D*** and determine how unlikely it would be to observe *D* if *H* were false.

In other words we are considering P(*D* | *not H*).

If this probability is sufficiently low, e.g 5% (a particular *p*-value) then we have a certain ‘level of confidence’ that we can ‘reject’ the ‘null hypothesis’, i.e. that *H* is false.

We say we reject the hull hypothesis at the 5% significance level.

Typical *p*-values are 5%,1%, 0.5%

Classical hypothesis testing and *p*-values

Note:
If we find P(*D* | *not H*) < 5%

Then this is the same as

P(*not D* | *not H*) > 95%

So: “we are at least 95% sure this data would *not* be observed if *H* is false”

But it does NOT mean we are at least 95% sure H is true.

We cannot conclude ***anything*** about P(*H*) without also knowing both:
the prior probability for *H* and
P(*D* | *H*)   (by Bayes’ Theorem)

Classical hypothesis testing and *p*-values

Example: Suppose ***H*** is the hypothesis: “This coin is biased”

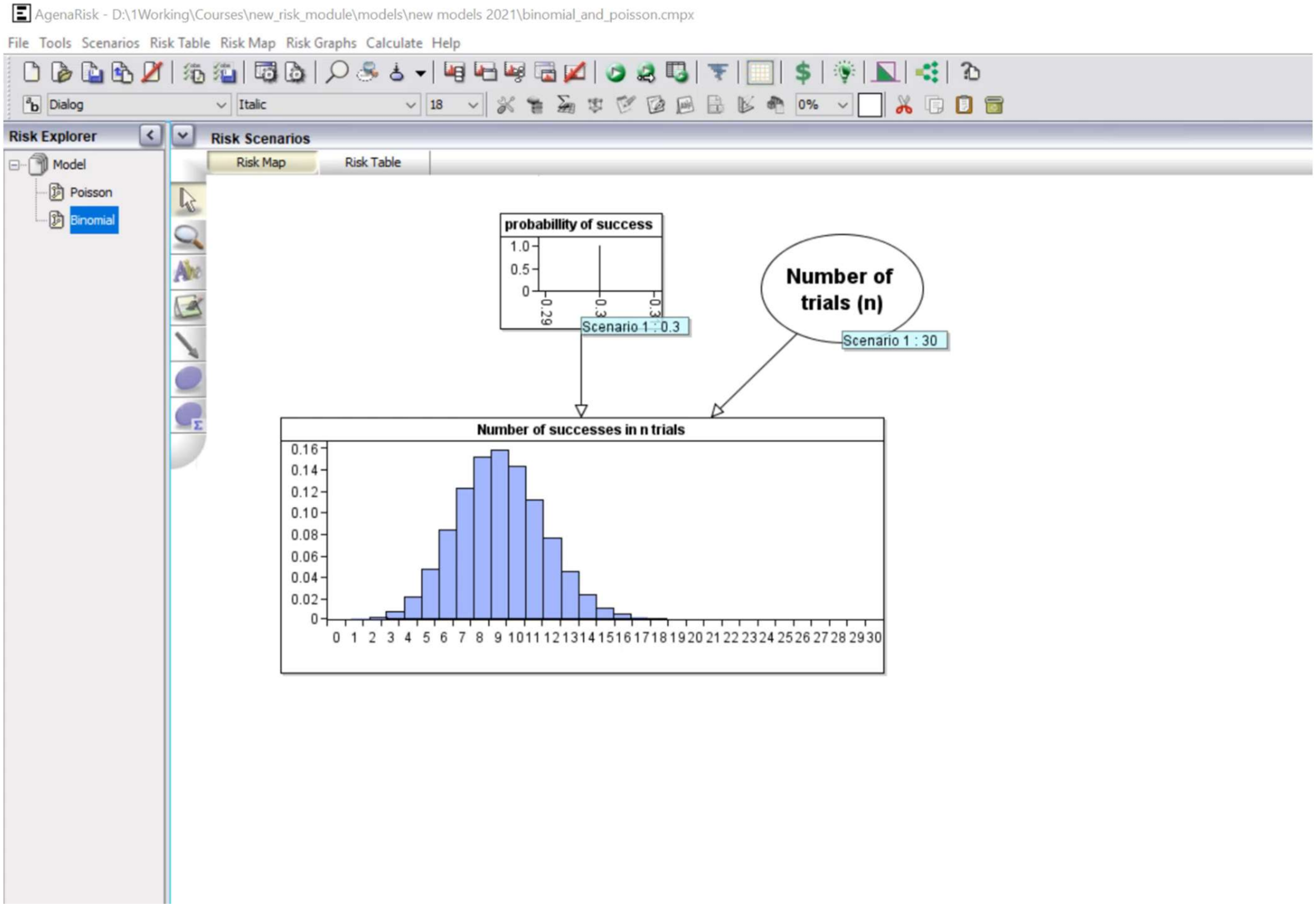
So the null hypothesis ***not H*** is
“This coin is unbiased” i.e.
there is a 50% chance it will land Heads up if I toss it

We want to test the null hypothesis at the 5% significance level.

We toss the coin 10 times and get 7 Heads. So the evidence D is “7 heads out of 10”)

Can we reject the null hypothesis?

….we need to calculate P(D | not H) and see if it is less than 5%



Scenario 1 : 0.3

Scenario 1 : 30

Classical hypothesis testing and *p*-values

Using Binomial distribution we calculate P(D | not H) = 11.7%

So we ***cannot*** reject the null hypothesis at the 5% level.

We do not have sufficient evidence.

So suppose we toss the coin 20 times and get 14 Heads.

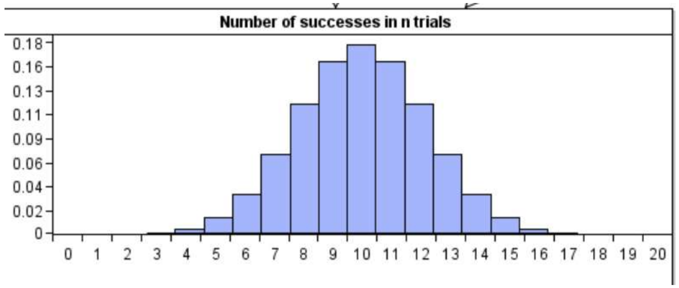
Can we reject the null hypothesis?

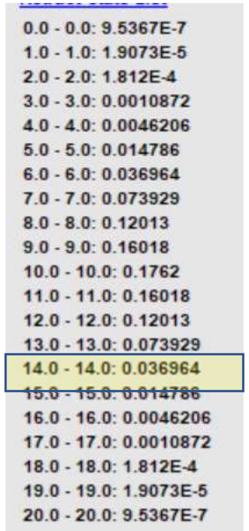
Yes we ***can*** at the 5% level because P(D | not H) = 3.69%

If we toss the coin 40 times and get 28 Heads we find P(D|not H) = 0.5% so we can reject the null hypothesis at the 1% level.

This is normally regarded as ‘highly significant’.

Binomial distribution: number of Heads in 20 tosses of fair coin





```
0.0 - 0.0: 9.5367E-7
1.0 - 1.0: 1.9073E-5
2.0 - 2.0: 1.812E-4
3.0 - 3.0: 0.0010872
4.0 - 4.0: 0.0046206
5.0 - 5.0: 0.014786
6.0 - 6.0: 0.036964
7.0 - 7.0: 0.073929
8.0 - 8.0: 0.12013
9.0 - 9.0: 0.16018
10.0 - 10.0: 0.1762
11.0 - 11.0: 0.16018
12.0 - 12.0: 0.12013
13.0 - 13.0: 0.073929
14.0 - 14.0: 0.036964
15.0 - 15.0: 0.014786
16.0 - 16.0: 0.0046206
17.0 - 17.0: 0.0010872
18.0 - 18.0: 1.812E-4
19.0 - 19.0: 1.9073E-5
20.0 - 20.0: 9.5367E-7
```

See model
“binomial_and_poisson.cmpx”

Classical hypothesis testing and *p*-values

Unfortunately there are multiple problems with this approach to hypothesis testing. Notably,

In most situations there is no simple way to calculate P(D | not H).

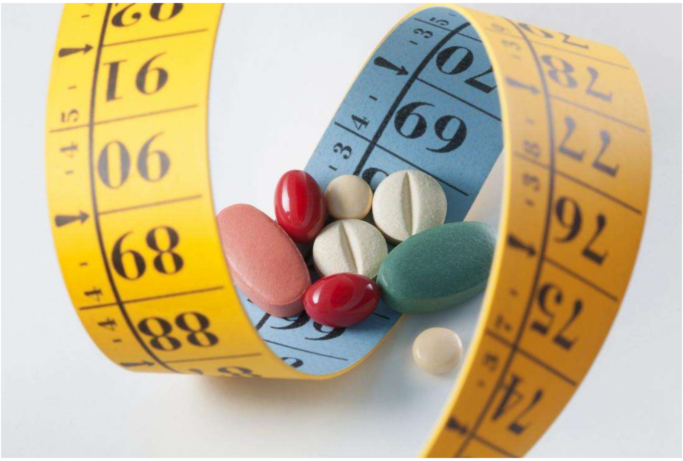
- For example, suppose the hypothesis is that “mean weight loss is at least 5lb if people take a particular diet pill”. The evidence will be in the form of numbers like 4, 10, 0, 6, -1, 0, 7, …. corresponding to subject’s weight loss.
- The standard classical approaches to estimating P(D | not H) is such cases make assumptions which are quite irrational and are poorly understand by practitioners who simply follow ‘rules’.

While setting an ‘agreed’ p-value in advance is supposed to stop people from arbitrarily claiming significance (“p-hacking”) there is no scientific justification for rejecting the null hypothesis at say 5% but not at 5.01%

p-values reward low variation more than magnitude of impact

Classical hypothesis testing and p-values

Example: Is our new weight loss drug "Precision" effective?



H: "People taking the drug lose weight over a 6-month period"
Null hypothesis (not H): "People taking the drug lose no weight over a 6-month period"

D: We observe 100 people using the drug. The average (mean) weight loss is 0.5 lb and the sample standard deviation 2.05. The *standard deviation of the mean* is then calculated as

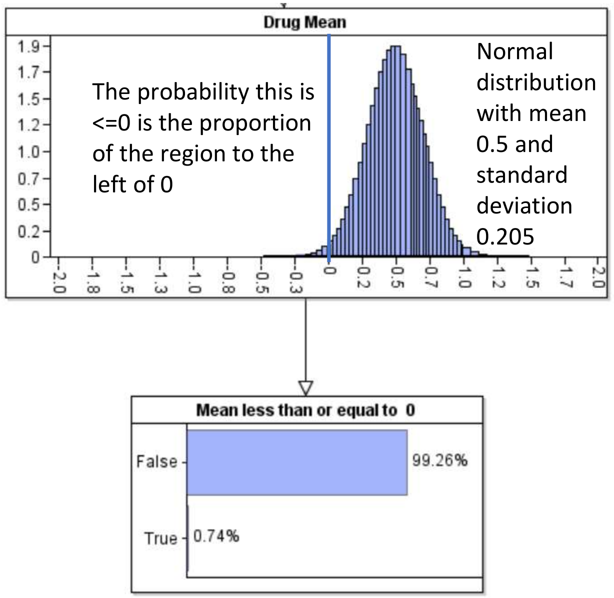
$$\frac{\text{sample standard deviation}}{\sqrt{\text{sample size}}} = \frac{2.05}{10} = 0.205$$

Providing (as in the case) that the sample size is at least 30, the 'classic' way to estimate P(D | not H) is to:

1. Assume the (true) *mean weight loss* has a Normal distribution whose mean is 0.5 and standard deviation 0.205
2. Then calculate the probability that this distribution is less than or equal to zero

Using standard tables, excel or AgenaRisk you can see in this case that the probability (which is also called the p-value) is 0.0074, i.e. 0.74%

As the p-value < 1% we reject the null hypothesis at the 1% p-value level – and hence 'accept' that there is 'significant' support for H.

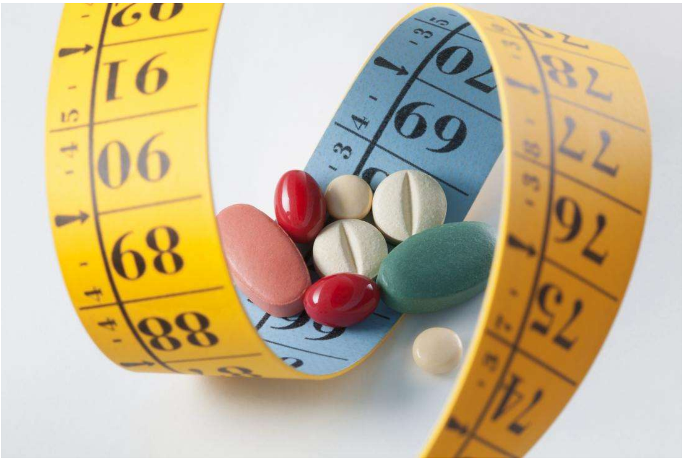


The probability this is <=0 is the proportion of the region to the left of 0

Normal distribution with mean 0.5 and standard deviation 0.205

See excel spreadsheet "oomph v precision" for raw data and calculations

Classical hypothesis testing and p-values



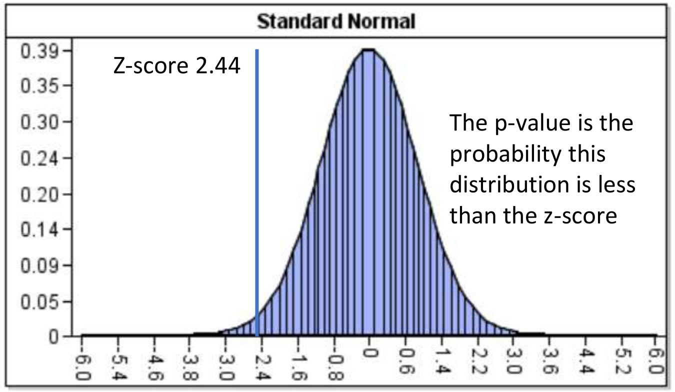
In the old days it was necessary to ‘transform’ the particular Normal distribution into a ‘standard Normal distribution’ (one with mean 0 and standard deviation 1) in order to calculate the p-value because – in the absence of computers – people relied on tables that had the standard normal distribution.

To do the transformation we calculate what is called the z-score:

$$z = \frac{\text{sample mean } - \text{null hypothesis mean}}{\text{standard deviation of mean}} = \frac{0.5-0}{0.205} = 2.44$$

This z-score is the distance from the mean of the ‘standard Normal distribution’ (one with mean 0 and standard deviation 1). The p-value – which is exactly equivalent to the p-value 0.0074 we previously calculated – is equal to the probability that the distribution is less than the z-score.

Tables of standardized z-scores show that any value above 2.326 has a probability less than 1% - so we can reject the null hypothesis at the 1% level.



Z-score 2.44

The p-value is the probability this distribution is less than the z-score

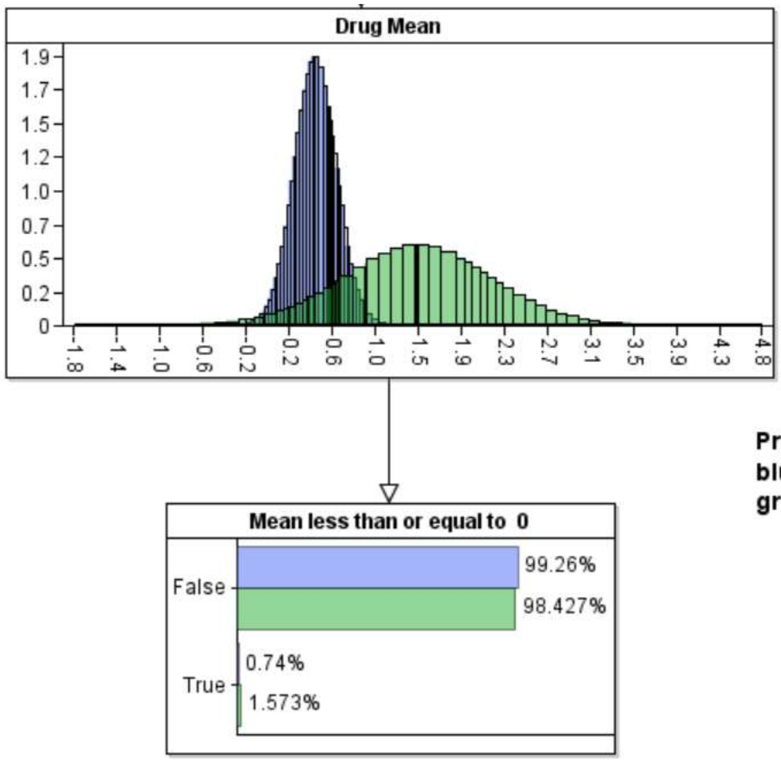
Which weight loss drug is best?

Precision*:* the mean weight loss for 100 subjects is 0.5 lbs with sample standard deviation 2.05 (so standard deviation of mean is 0.205)

Z-score 2.44, p-value 0.0074 (i.e. 0.74%). Null hypothesis rejected at 1%

Oomph*:* the mean weight loss for 100 subjects is 1.5 lbs with sample standard deviation 6.97 (so standard deviation of mean is 0.697)

Z-score 2.15, p-value 0.016 (i.e. 1.6%). Null hypothesis NOT rejected at 1%



Precision is blue, Oomph is green

p-values reward low variation more than magnitude of impact

Misinterpreting results of hypothesis tests

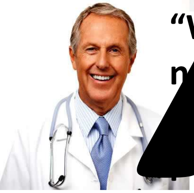
“Null hypothesis” H: drug X provides no patient benefit

Get data D on patients using drug X

A lot of good outcomes. Less than 5% chance of seeing so many if the drug was useless

We can ‘reject the null hypothesis’ with p-value 0.05 (i.e. 5%)

But what does that really mean?



“W 95% sure the n hesis is false, 95% e the drug ovides efit”

The probability of D given H is 5%.

This does not mean the probability of H given D is 5%

Example: A low p-value – but the null hypothesis is *certainly* true

See video “A simple example demonstrating the limitations of p-values for hypothesis testing” https://youtu.be/vk0rKIaGQBs

Suppose we add this double-headed coin to a bag containing 999 fair coins

Select a coin at random and toss it 100 times to test the hypothesis H “Coin is fair (i.e. not double-headed)”

Set p-value 0.05 in advance

Suppose we get 55 Heads out of 100.

Then P(E | H) = 0.0485 = 4.85%

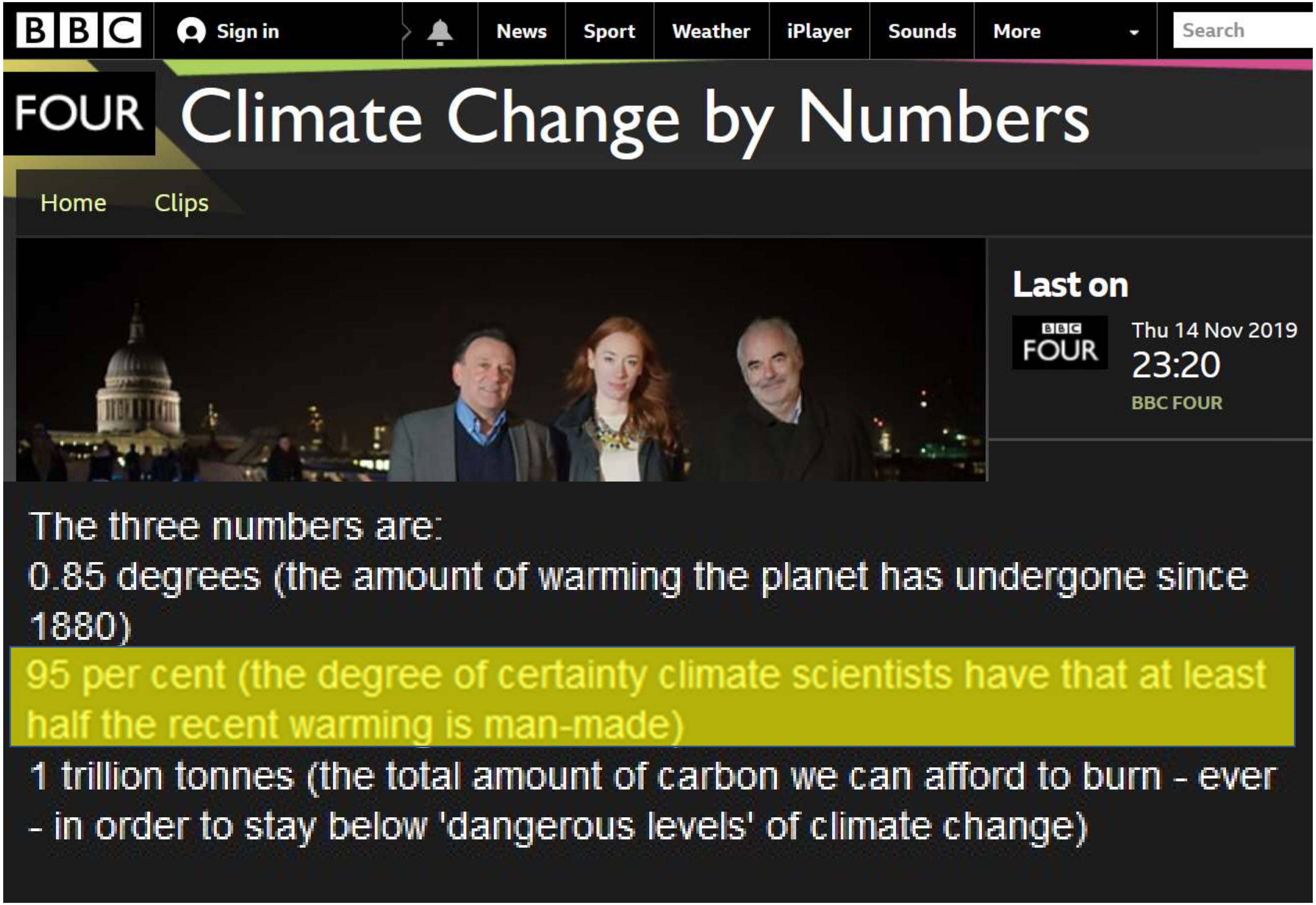
So we can reject H at the 5% level. Are we 95% certain that the coin is *not* fair?

No, in fact it is 100% certain the coin *is* fair.

In fact, for a fair coin, there is a probability 0.184, i.e. 18.4% chance of getting at least 55 Heads in 100 tosses.

So there is an 18.4% chance that the p-value test will wrongly ‘conclude’ that the coin is double-headed.

See model “coin_toss_hypothesis”. Use it to show that if we get 61 Heads we can reject at H at the 1% level



Climate Change by Numbers

Last on

BBC FOUR

Thu 14 Nov 2019
23:20
BBC FOUR

The three numbers are:

0.85 degrees (the amount of warming the planet has undergone since 1880)

95 per cent (the degree of certainty climate scientists have that at least half the recent warming is man-made)

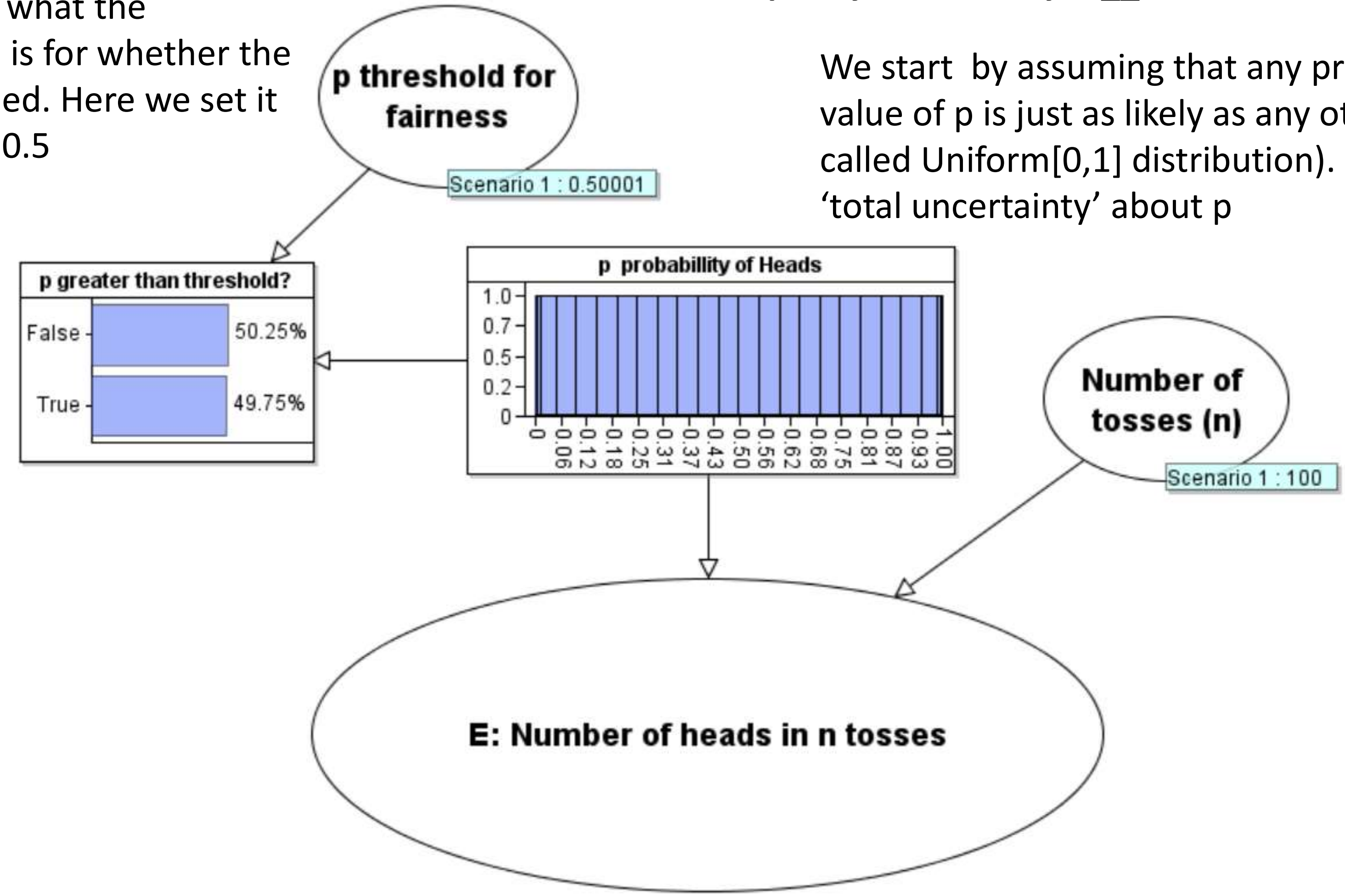
1 trillion tonnes (the total amount of carbon we can afford to burn - ever - in order to stay below 'dangerous levels' of climate change)

Simple Bayesian hypothesis testing

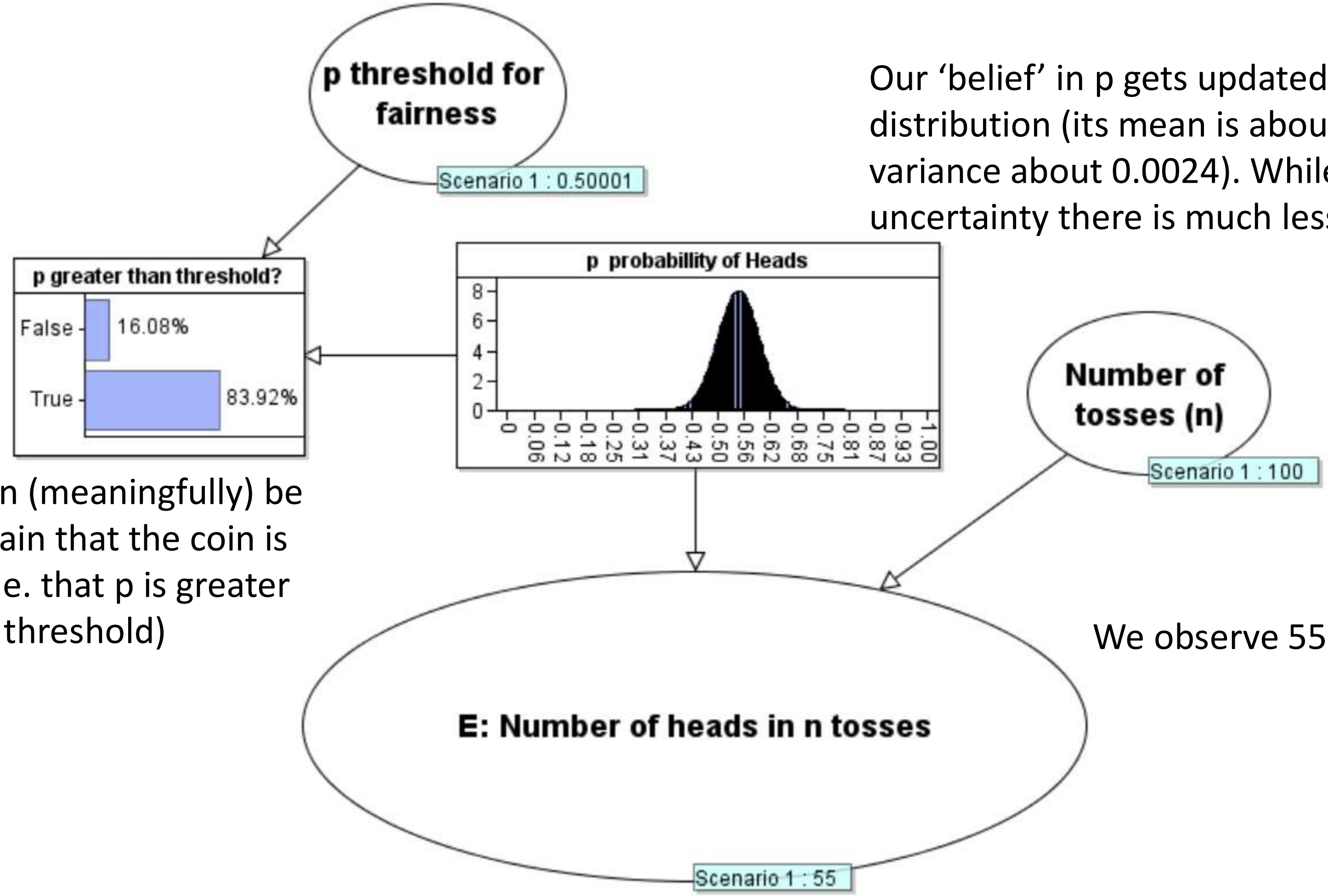
See video “A simple example of Bayesian hypothesis testing” https://youtu.be/s4yCu__18Jo

We decide what the ‘threshold’ is for whether the coin is biased. Here we set it just above 0.5

We start by assuming that any probability value of p is just as likely as any other (the so-called Uniform[0,1] distribution). So we have ‘total uncertainty’ about p



Simple Bayesian hypothesis testing

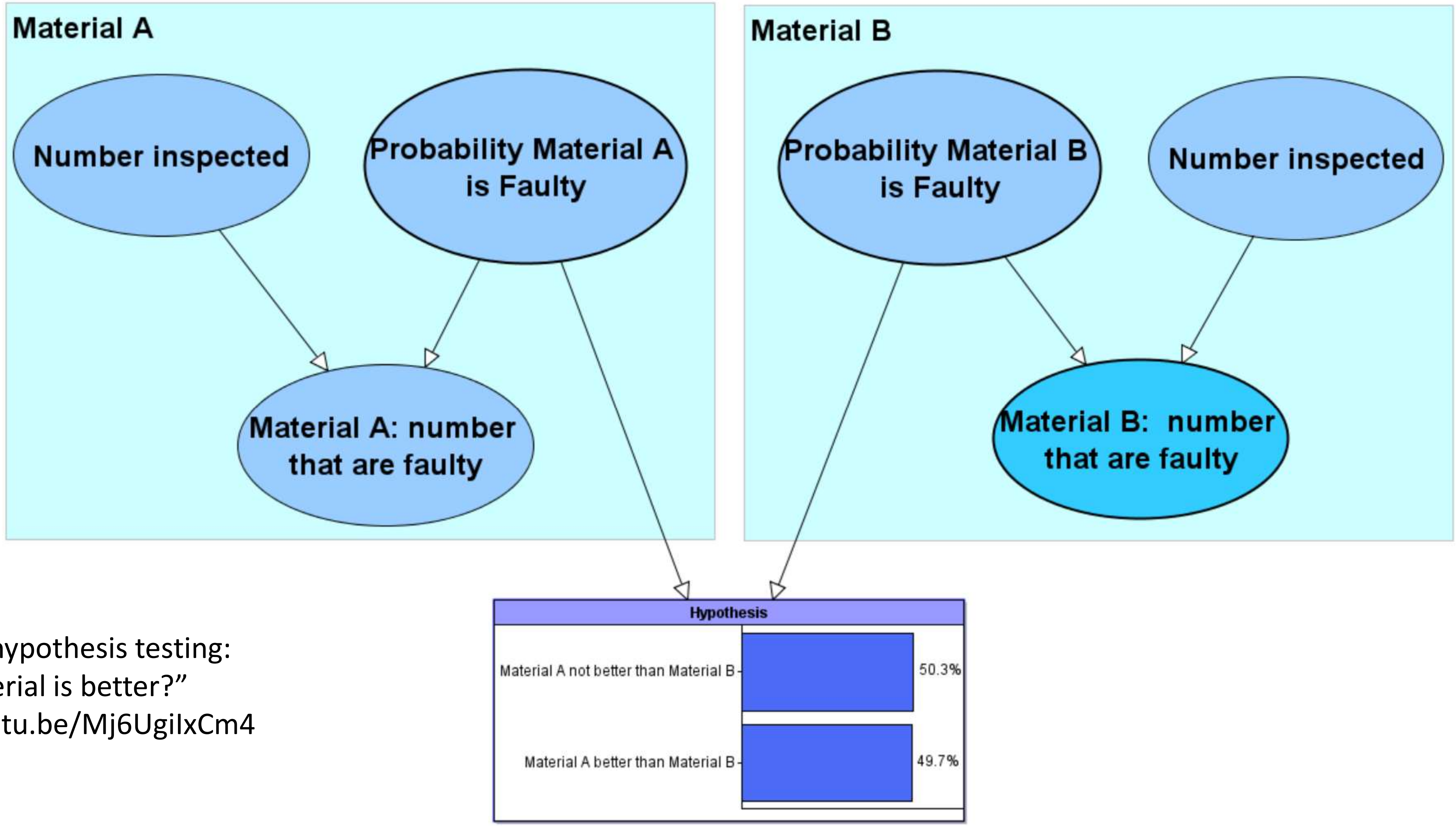


Our ‘belief’ in p gets updated to this distribution (its mean is about 0.55 and variance about 0.0024). While there is still uncertainty there is much less than before

So we can (meaningfully) be 84% certain that the coin is biased (i.e. that p is greater than the threshold)

We observe 55 heads in 100 tosses

Testing the difference between two ‘populations’


Material A
Number inspected
Probability Material A is Faulty
Material A: number that are faulty
Material B
Probability Material B is Faulty
Number inspected
Material B: number that are faulty
Hypothesis
Material A not better than Material B
50.3%
Material A better than Material B
49.7%

See video:
“Bayesian hypothesis testing:
which material is better?”
https://youtu.be/Mj6UgiIxCm4

Which – if any – drug should a patient with this disease take?

For patients with a particular medical condition the mortality rate is known to be 25% (i.e. 25% of patients die within a specified time).

A new treatment which can be properly demonstrated to reduce the mortality rate will be recommended to patients with this condition.

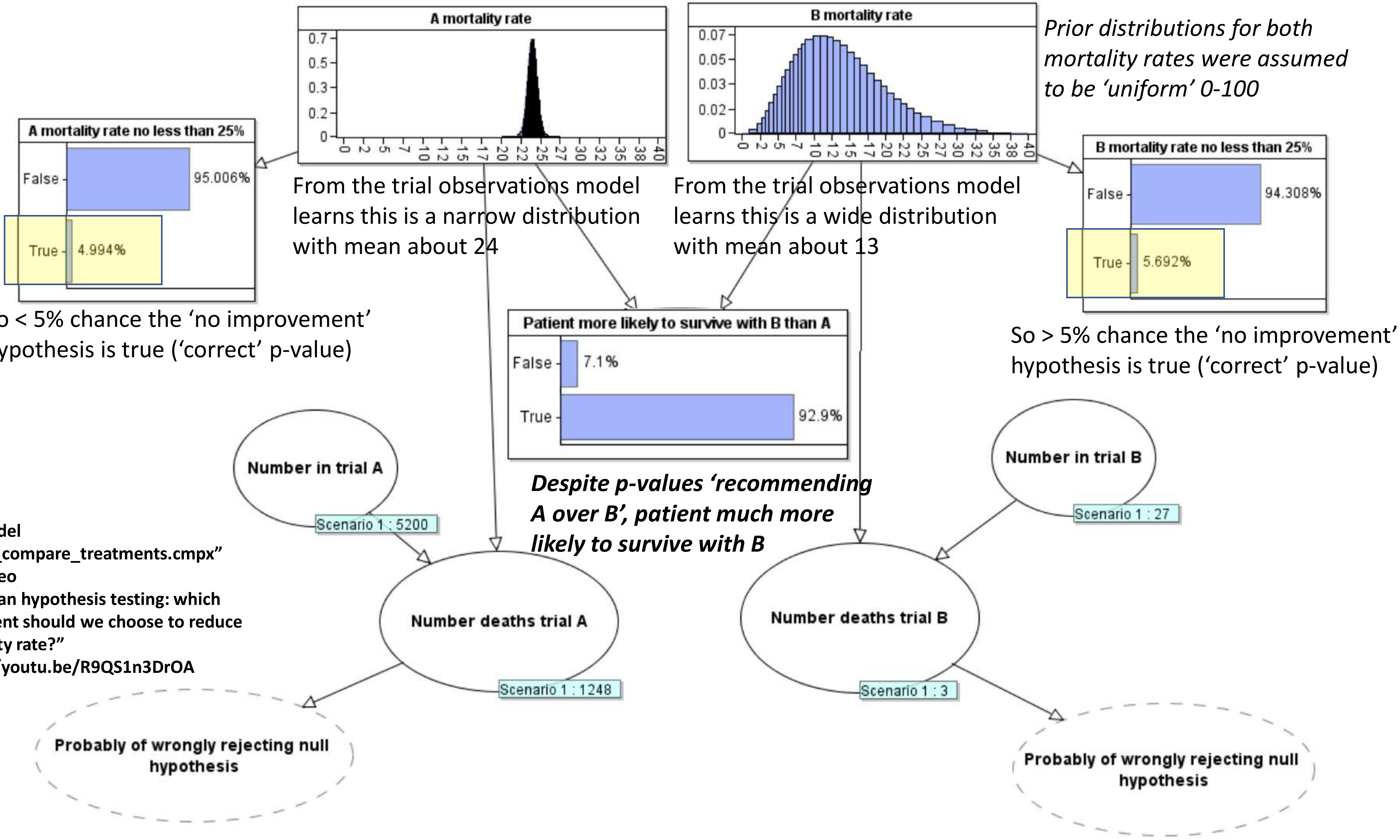
Two treatments A and B have been trialled and tested against the ‘null hypothesis’ of ‘no improvement’ on the 25% mortality rate.

- Treatment A is based on a large trial of 5,200 patients of whom 1,248 die, so the observed mortality rate is 24%
- Treatment B is based on small trial of 27 patients of whom 3 die, so the observed mortality rate is just over 11%

Which treatment should be chosen?

- A
- B
- Both A and B
- Neither A nor B

Thanks to Karim Brohi for proposing this question



Prior distributions for both mortality rates were assumed to be 'uniform' 0-100

From the trial observations model learns this is a narrow distribution with mean about 24

From the trial observations model learns this is a wide distribution with mean about 13

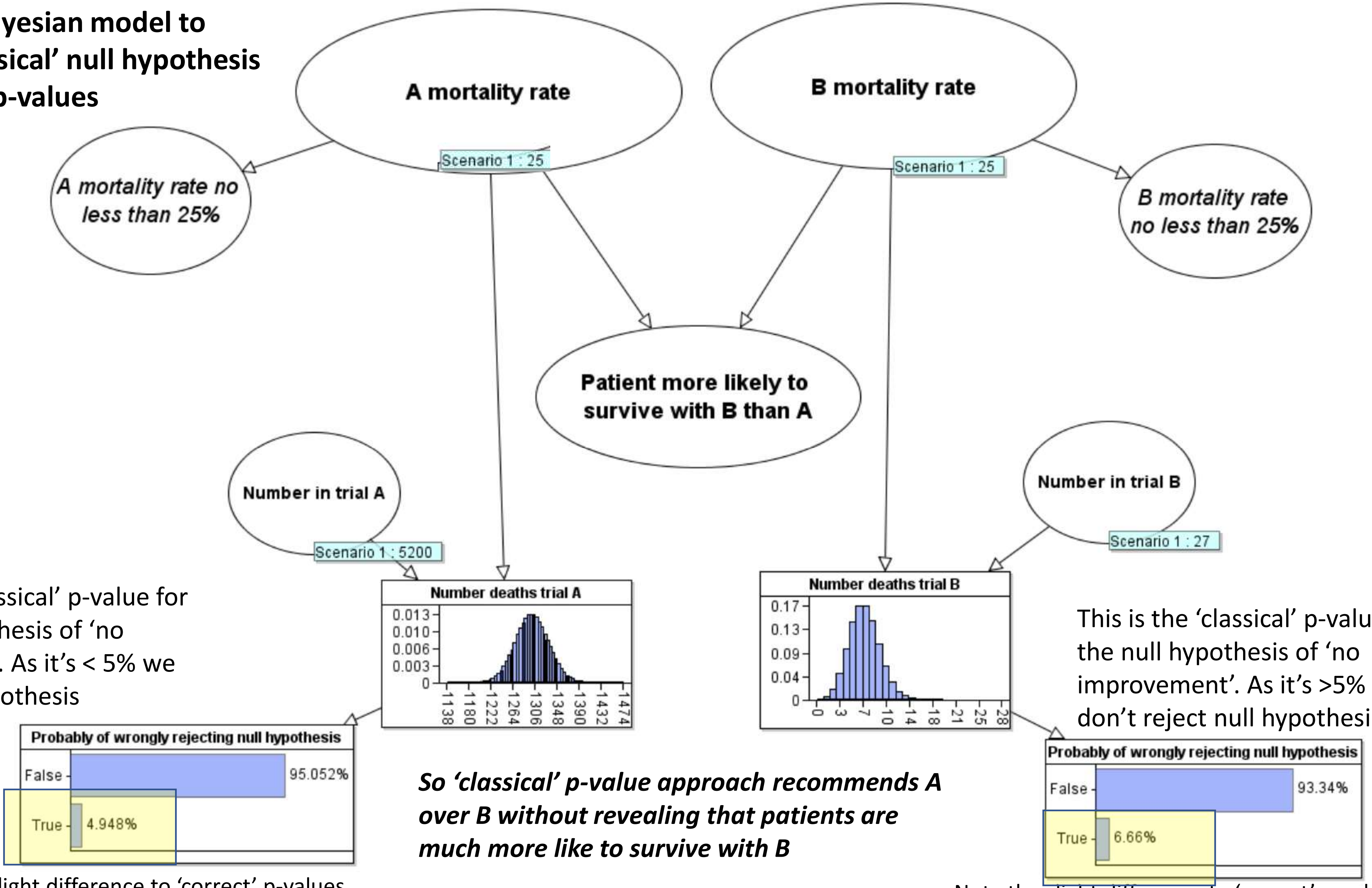
So < 5% chance the 'no improvement' hypothesis is true ('correct' p-value)

So > 5% chance the 'no improvement' hypothesis is true ('correct' p-value)

Despite p-values 'recommending A over B', patient much more likely to survive with B

See model "karim_compare_treatments.cmpx" and video "Bayesian hypothesis testing: which treatment should we choose to reduce mortality rate?" https://youtu.be/R9QS1n3DrOA

Using the Bayesian model to explain 'classical' null hypothesis testing and p-values



This is the 'classical' p-value for the null hypothesis of 'no improvement'. As it's < 5% we reject null hypothesis

Note the slight difference to 'correct' p-values

So 'classical' p-value approach recommends A over B without revealing that patients are much more like to survive with B

This is the 'classical' p-value for the null hypothesis of 'no improvement'. As it's >5% we don't reject null hypothesis

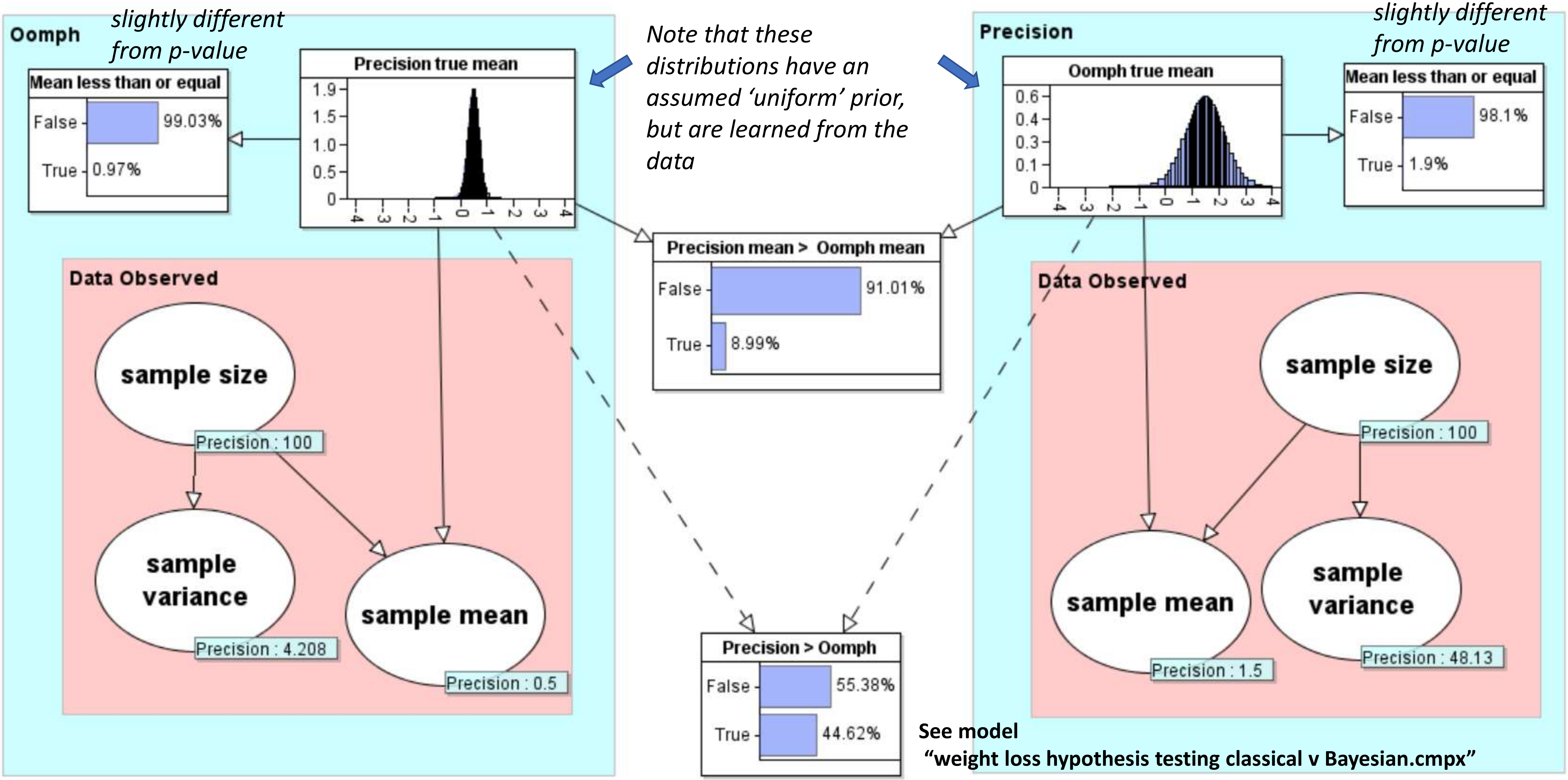
Note the slight difference to 'correct' p-values

Bayesian solution of diet drug problem

Result here is slightly different from p-value

Note that these distributions have an assumed 'uniform' prior, but are learned from the data

Result here is slightly different from p-value



See model "weight loss hypothesis testing classical v Bayesian.cmpx"

Key points

- Most claims about critical risks based on empirical, experimental studies are misinterpreted.
- This is because the results of classical statistical null hypothesis testing are poorly presented
- There are fundamental problems and limitations of classical null hypothesis testing
- In particular p-values are misunderstood and misleading
- Bayesian hypothesis testing addresses the fundamental limitations of classical null hypothesis testing