ECS647U / ECS773P

Bayesian Decision and Risk (BDRA)

Semester 2, 2021

Coursework 1 – Tricky Magician

Mariya Pavlova

Queen Mary University of London

MSc Artificial Intelligence

Student ID 170703132

Part 1: Background

1. **Specify the Binomial distribution and describe under what circumstances its use is appropriate. [4]**

The Binomial distribution could be defined as the probability of a success of failure outcome of an event that is repeated multiple times. The binomial distribution always has two possible outcomes.

For example:

- Taking a test has two possible outcomes: pass or fail.
- Tossing a coin has two possible outcomes: heads or tail.

The Binomial distribution is defined by two parameters *n* (number of trials, for example) and *p* (the success probability). Below is the general formula for the Binomial distribution.

$$\frac{n!}{r! \times (n-r)!} \times p^r (1-p)^{n-r}$$

Fig. 1: General Formula for the Binomial Distribution (1)

- Where the first variable *n* stands for the number of times the experiment runs.
- *p*: represents the probability of one specific outcome.

More formally, the binomial distribution with parameters *n* and *p* is the discrete probability distribution of the number of successes in a sequence of *n* independent experiments. Each of these experiments is asking a yes–no question, and each can be either success (with probability p) or failure (with probability 1 − p).

Binomial distribution must also meet the following criteria:

- Fixed number of trials: there must be a clearly defined number of trials that do not vary. This number cannot be altered at any point and each trial must be performed the same way as all of the others.
- Independent trials: each trial should have no effect on any of the others.
- Two different classifications: success or failure.
- The probability of success stays the same for all trials.

Given the characteristics of the Binomial distribution, we can conclude that all the above criteria must be present to use the Binomial probability.

Binomial distribution could be used in cases where the distribution of data is binary and from a finite sample. Thus, it gives the probability of getting *r* events out of *n* trials. A typical example of using Binomial probability is tossing a coin. If we toss it once, the probability of getting heads is 50%. If we toss a coin many times, the probability of getting heads gets closer and closer to 100%.

2. **<u>Specify the probability density function for the Beta distribution and state how the expected value is calculated? [4]</u>**

The Beta distribution belongs to the family of continuous probability distributions. It is defined on the interval [0, 1] because it models a probability. It has two positive shape parameters: α and β. They appear as exponents of the random variable and control the shape of the distribution. When building a BN, instead of using normal assumptions, we could use a Beta distribution with parameters alpha and beta (1). In other words, α and β would substitute the mean and variance in the model.

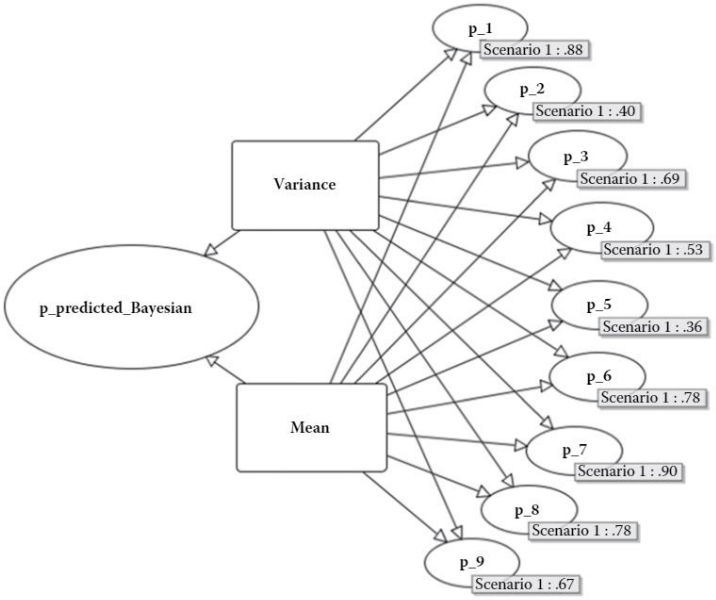

Fig. 2: Structure of parameter learning model with normal assumptions.
We can use the model for Beta distribution too(1).

The Beta distribution is the conjugate prior for the Bernoulli, binomial, negative binomial and geometric distributions in Bayesian inference (1). It is important because it helps us avoid expensive numerical computations involved in Bayesian Inference. I.e computing a posterior using a conjugate prior.

The probability density function for the Beta distribution is presented below (1).

E.2.1 Beta

Probability function: $P(X) = \frac{(1-x)^{\beta-1} x^{\alpha-1}}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} (1-x)^{\beta-1} x^{\alpha-1}$
Domain: $0 \le X \le 1$
Parameter domain(s): $\alpha > 0$, $\beta > 0$
Mean: $E(X) = \frac{\alpha}{\alpha+\beta}$
Variance: $V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Note: The domain of the Beta distribution can be extended to any finite range in the region $L \le X \le U$.
Example: Beta(3, 7, 0, 10)

The probability density function of the beta distribution, for 0 ≤ x ≤ 1, and shape parameters α, β > 0, is a power function of the variable x and of its reflection (1 − x).

Where Γ(z) is the gamma function. The beta function is a normalisation constant. It ensures that the total probability is 1. In the below equations x is an observed value that occurred — of a random process X.

1/B(α,β) in the density function is a normalising constant to make the function integrates to 1. Then, the terms in the numerator, x^(α-1) * (1-x)^(β-1), are familiar. This is binomial distribution. The difference is that the Binomial distribution models the number of successes while the Beta distribution models the probability of success. In other words, in Beta the probability is a random variable.

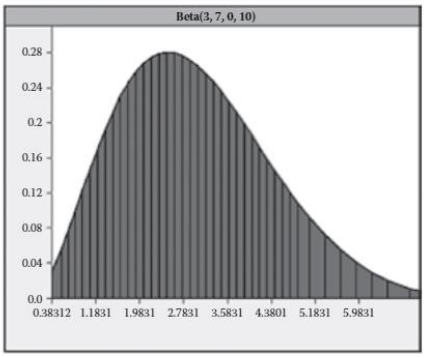

Fig. 3: Beta(3,7,0,10) distribution example (1)

3. <u>What role do the parameters play in determining the shape of the distribution, and describe how it might be used to express a prior belief about the chance of a fair or unfair coin? [6]</u>

In terms of α and β in the probability density function, α-1 could be described as the number of successes, and β-1 as the number of failures. This is analogous to the terms in Binomial distribution. We set α and β according to how big we think the probability of success or failure

is. As α becomes larger (more successful events), most of the probability distribution will shift towards the right. On the other hand, an increase in β moves the distribution towards the left. This means more failures. Also, the distribution will narrow if both α and β increase because the outcome certainty increases.

In general, when we change the prior parameters, the posterior distribution changes accordingly. It illustrates the strength of the prior on the posterior results. Generally, when the prior is strong and the sample is low, the prior dominates. When the prior is weak and the sample is high, the data dominates. The range of the distribution gets smaller as the sample size increases.

It can be demonstrated with an example of the chance of a fair or unfair coin in a coin-tossing experiment. First, we shall determine the prior distributions of α and β. We can do that by using fixed values that represent ignorance, strong, or biased assumptions about the probability of a coin coming up head (i.e being unfair). The we shall add observations to learn the posterior parameters for p. It allows us to calculate the posterior under several different conditions, combining different priors with different amounts of proof. Our model will contain 9 different scenarios. The model and examples were taken from the study book (1).

$P(\alpha = 1, \beta = 1 \mid \text{ignorant})$, $P(\alpha = 10, \beta = 10 \mid \text{strong})$, $P(\alpha = 9, \beta = 1 \mid \text{biased})$

n = 100 trials (coin flips for each belief)

- **Ignorant prior belief**: mean p(coin=heads) = 0.5 for a fair coin. Let us have two experiments: n trials resulting in 5 heads, and another n trials resulting in 50 heads. Over many flips we get a result centred around 0.5.
- **Strong prior belief**: mean p(coin=heads) = 0.5 for a fair coin. Let us have the same experiments as above with n resulting in 5 heads, and n resulting in 50 heads. Over many flips we get a result centred around 0.5 but with much stronger belief than with ignorant prior.
- **Biased belief** (unfair coin): now α and β will have a mean of p(coin=heads) = 0.9. Let us have the first n trial resulting in 5 heads, and the second in 50 heads. Over many runs we get a result that shifts progressively towards 0.5.

Let us illustrate the experiment in Fig.4 below.

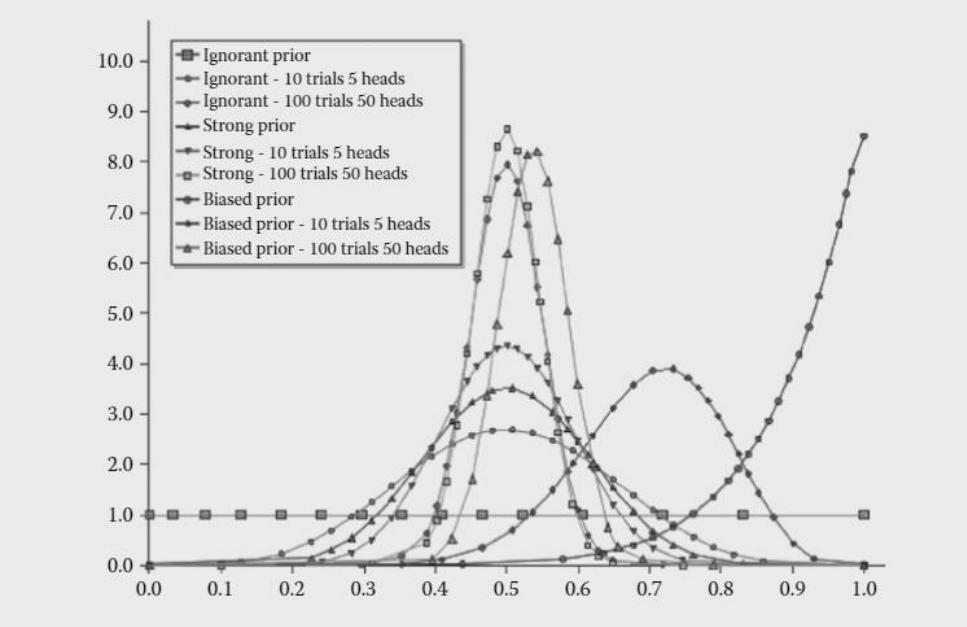

Fig. 4: Posterior marginal distributions for Beta Binomial with fixed values and different sets of observations (1)

We can make the following observations:

- the ignorant prior is completely flat.
- the strong prior is centred on a half, i.e there is a predisposition to favour values closer to this region.
- the biased prior favours values closer to one.

We can also derive from the experiments that as we add data the parameter value learned changes. The more data we add, the more dramatic the change. For example, adding 50 heads from 100 trials changes the biased prior to a posterior almost centred on 0.5. Thus, in this case, the data (the likelihood) could be said to overthrow the prior. We can also notice that the ignorant prior is updated by whatever data we provide to it such that the posterior is simply equivalent to the data (likelihood). In other words, the data could be said to speak for themselves.

4. **Use a Beta-binomial formulation to specify the posterior distribution of P(X | n, p) where X is the number of heads, n is the number of flips and p is the probability of the coin coming up heads in each flip. [6]**

The Beta-binomial distribution is the binomial distribution in which the probability of success at each of n trials is not fixed but randomly drawn from a beta distribution.

We have the following prior beliefs:

- Binomial distribution Bin(n,p) with n known and p unknown

- Prior belief about p is beta(α,β)

Next, we observe x success in n trials. The Bayes' rule implies that the new belief about the probability density of p is also the beta distribution. It has different parameters though. In mathematical terms, it is expressed in the following way:

$p|x \sim beta(\alpha+x, \beta+n-x)$,(equation 1)

In other words, the posterior distribution appears to be in the same family of probability density functions as the prior belief but with new parameter values. These have been updated to reflect the new observations from the data, i.e what the model learned from the data. This is called conjugacy. The compatibility between the Binomial and Beta distributions is due exactly to conjugacy. Mathematically, the resulting posterior distribution has the same conjugal form as the prior used, thus ensuring analytical tractability (1). For example, a Beta prior distribution multiplied by a Binomial likelihood results in a Beta posterior distribution (1). This Beta-Binomial pairing is explained in mathematical terms in Fig. 5:

We have already encountered the Binomial distribution. The likelihood is

$$P(X \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

We can model p as the Beta function where (α, β) are the number of successes and failures respectively in a finite number of trials. This is thus identical to the idea of Bernoulli successes and failures in the Binomial model. The prior model is

$$P(p \mid \alpha, \beta) = Beta(\alpha, \beta)$$

The neat thing about the Beta distribution is that the conditional posterior distribution is still a Beta distribution (proven by way of Bayes' theorem):

$$P(p \mid X, \alpha, \beta) = Beta(\alpha + x, \beta + n - x)$$

Fig. 5: Beta-Binomial Hierarchical Model Formulations

We can show mathematically how Beta-Binomial conjugates, i.e how we get the posterior (2).

Recall the discrete of the Bayes rule:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

It should be noted that this formula does not apply to continuous random variables. For example, the P which follows a beta distribution. It because the denominator sums over all possible values of the random variable (2).

It does have a finite range though. This means, it can take any value between 0 and 1. Thus, integration can be performed. Now we can rewrite the Bayes rule as follows:

$$\pi^*(p|x) = \frac{P(x|p)\pi(p)}{\int_0^1 P(x|p)\pi(p)dp}.$$

This is like the discrete form because, just like in summation, the integral in the denominator will also be equal to a constant. This constant ensures that the posterior density function equals 1 (2).

The first term in the numerator, P(x|p), is our likelihood. The second term, π(p), is the probability density function that reflects the prior belief about p. In Beta-Binomial, we have: P(x|p)=Bin(n,p) and π(p)=beta(α,β)

When we plug in these distributions, we get:

$$\begin{aligned}\pi^*(p|x) &= \frac{1}{\text{some number}} \times P(x|p)\pi(p) \\ &= \frac{1}{\text{some number}}[\binom{n}{x} p^x(1-p)^{n-x}][\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}] \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \times p^{\alpha+x-1}(1-p)^{\beta+n-x-1}\end{aligned}$$

Let $\alpha^* = \alpha + x$ and $\beta^* = \beta + n - x$, and we get

$$\pi^*(p|x) = \text{beta}(\alpha^*, \beta^*) = \text{beta}(\alpha + x, \beta + n - x),$$

as the posterior formula in (Equation 1).

We can recognise the posterior distribution from the numerator $p^{\alpha+x-1}$ and $(1-p)^{\beta+n-x-1}$. Everything else are constants that must ensure that the area under the curve is between 0 and 1 equals 1. Thus, these values have to take the values of beta, which has parameters α+x and β+n− x (2).

In the end, it is also worth mentioning that despite the great advantages of conjugacy, it does not come at zero cost (1). The disadvantage is that when we build models, we often worry whether prior distributions are provably noninformative (1). Thus, it is important to know when to use conjugate priors in order to use them correctly.

Part 2: Basic Analysis

2.1 You have been handed a coin that has indeed been shaved by a magician, but you are indifferent to which side has been shaved. Model two hypothesised prior beliefs about the coin coming up heads on a given spin.

a) shaven on heads
b) shaven on tails

Here we shall use hyperparameters (α, β) to determine our uncertainty. We can think of the parameter α as representing the number of successes and β as representing the number of failures. In other words, the sample size is (α+β) and the success rate (or population mean) is α/α+β. The sample size and mean are estimable rather than purely mathematical abstractions. Thus, they are more natural hyper-parameters than α and β(1). Moreover, by making these parents of α and β we simply define α and β by the deterministic functions in Fig.6:

$$\alpha = mean \times sample\ size \quad since\ \alpha = \left(\frac{\alpha}{\alpha+\beta}\right) \times (\alpha+\beta)$$

$$\beta = (1-mean) \times sample\ size \quad since\ \beta = \left(1-\frac{\alpha}{\alpha+\beta}\right) \times (\alpha+\beta)$$

Fig. 6: Beta-Binomial Hyperparameter Specification

Now we shall create a model in AgenaRisk. Fig.7 shows the output. The AgenaRisk model file is attached to the pdf. It is called assignment_1_question 2.1.

Model parameters:

- P(α = 20 heads, β = 10 tails| shaven on heads)
- P(α = 10 heads, β = 20 tails| shaven on tails)

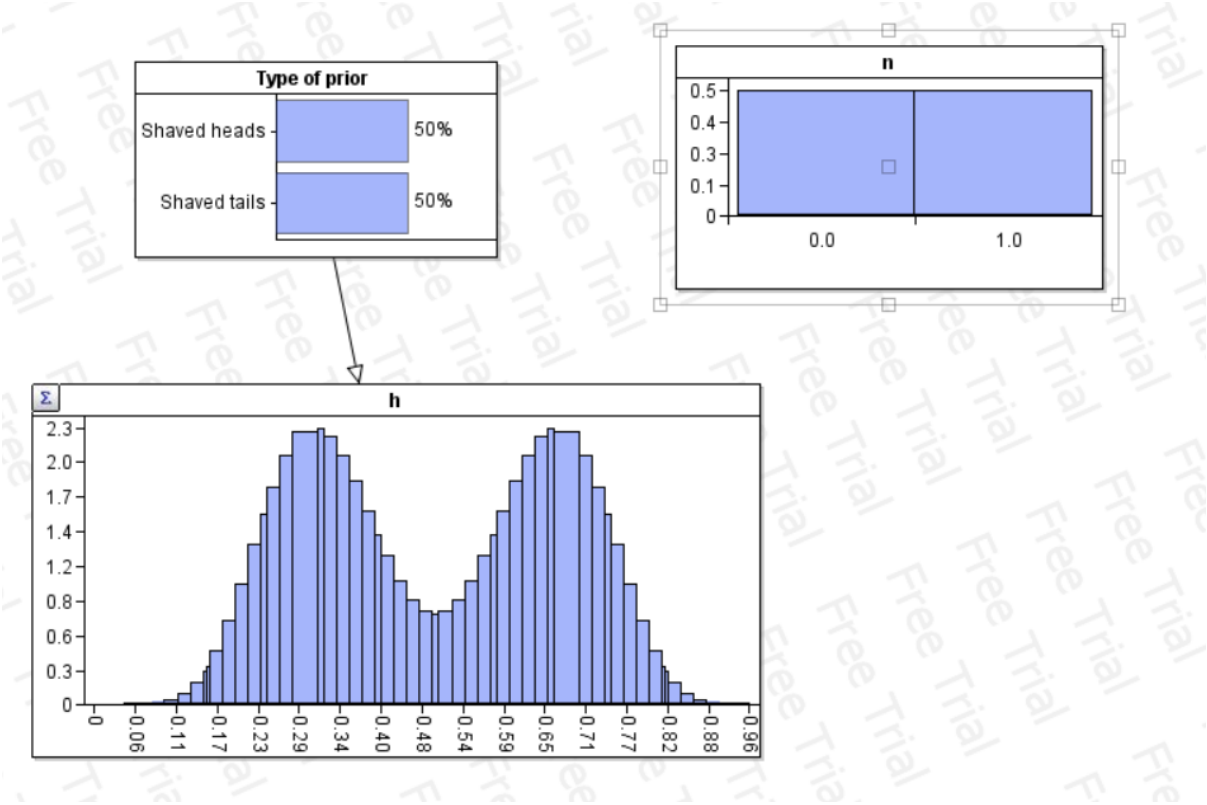

Fig. 7: AgenaRisk Model: coin shaved on heads or tails

We need to have sensible priors for sample size and mean. For the mean a reasonable prior distribution is Uniform [0,1] since it does not favour any value in this range (1).

In the model we can see that shaved heads and shaved tails are equally probable. This is because we do not have any observations yet. In other words, we are saying that our nodes are independent and that we have a strong prior and weak sample.

2. You are allowed 25 spins of the shaved coin and you observe the following results:

H H H T H H T H H H T T H T H H T H H H H T H T H

Carefully describe and justify how you used your model and calculate the posterior probability of each hypothesis.

We shall now add observations to our model and use the n number of trials node. The AgenaRisk model is shown below in Fig.8. It is also provided in the AgenaRisk file assignment_1_question 2.1.

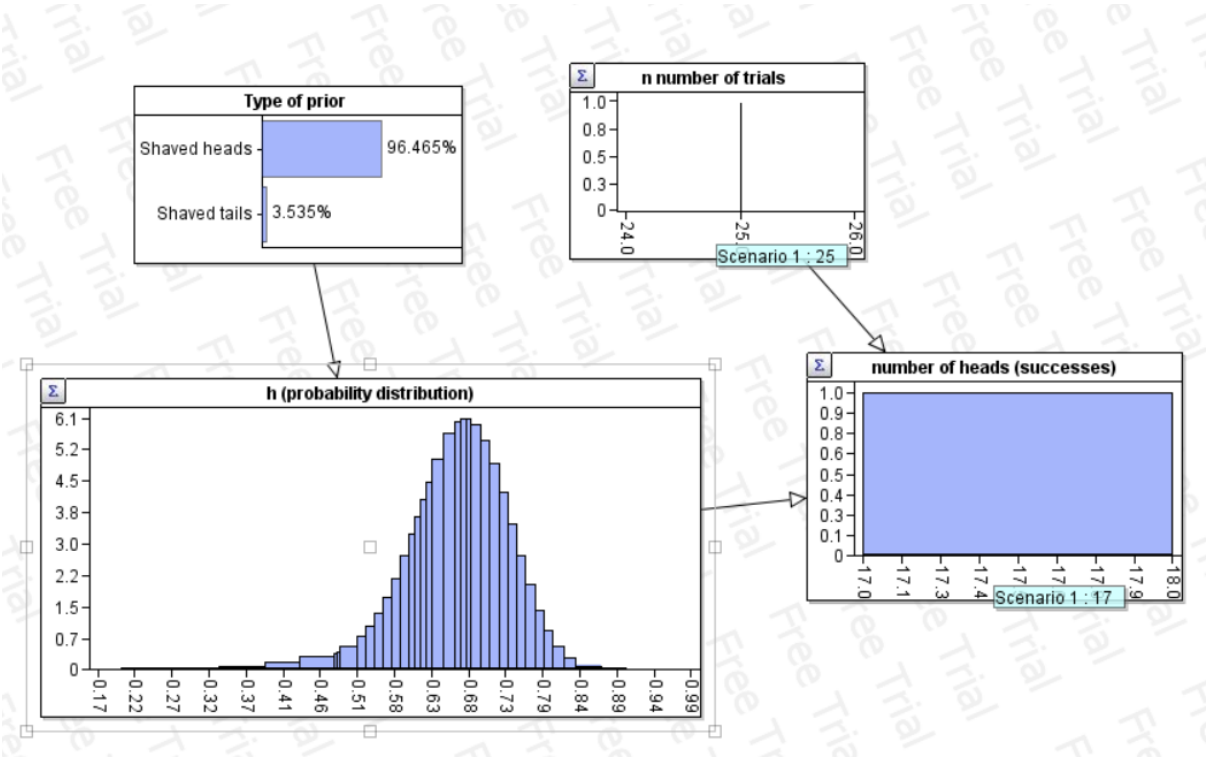

Fig. 8: AgenaRisk Model: coin biased towards heads

We have added the observations in the node called “Number of successes (heads)”. This has subsequently updated our posterior probability. It is now right skewed which means the coin is biased towards heads. As described in question 1, in this case the data speaks for itself. However, we do not have many trials. Thus, the probability distribution indicates we have a biased coin towards heads. The more observations we add, the narrower the distribution will get around 0.68. This would mean a very strong probability.

Part 3: Advanced Analysis

Repeat part 2 of the analysis but where you are told that there is a 2:1 chance that the magician has shaved the head of the coin.

3.1 Describe how this information about the magician's bias would change your prior beliefs about each hypothesis. [5]

The new information we received means that there is 2/3 chance that we would get heads, and that there is 1/3 chance that we would get tails. As described in part 1 of the assignment, this means that our prior dominates as there is no data. To illustrate, it would mean that the distribution would be more skewed to the right.

3.2 Carefully describe and justify your model and show a graph of the prior distribution for p, as a probability density function or histogram, for the chance of heads on any future spin under each hypothesis.

We can illustrate the above statement in AgenaRisk. We can use the model we created in part 2.1 of the assignment. The only change we shall make is adding the 2:1 chance that the magician has shaved the head of the coin. We will add it to our prior node and express it as a percentage. We can do this by changing our prior to reflect 2:1. Please see a screenshot of the model below in Fig.

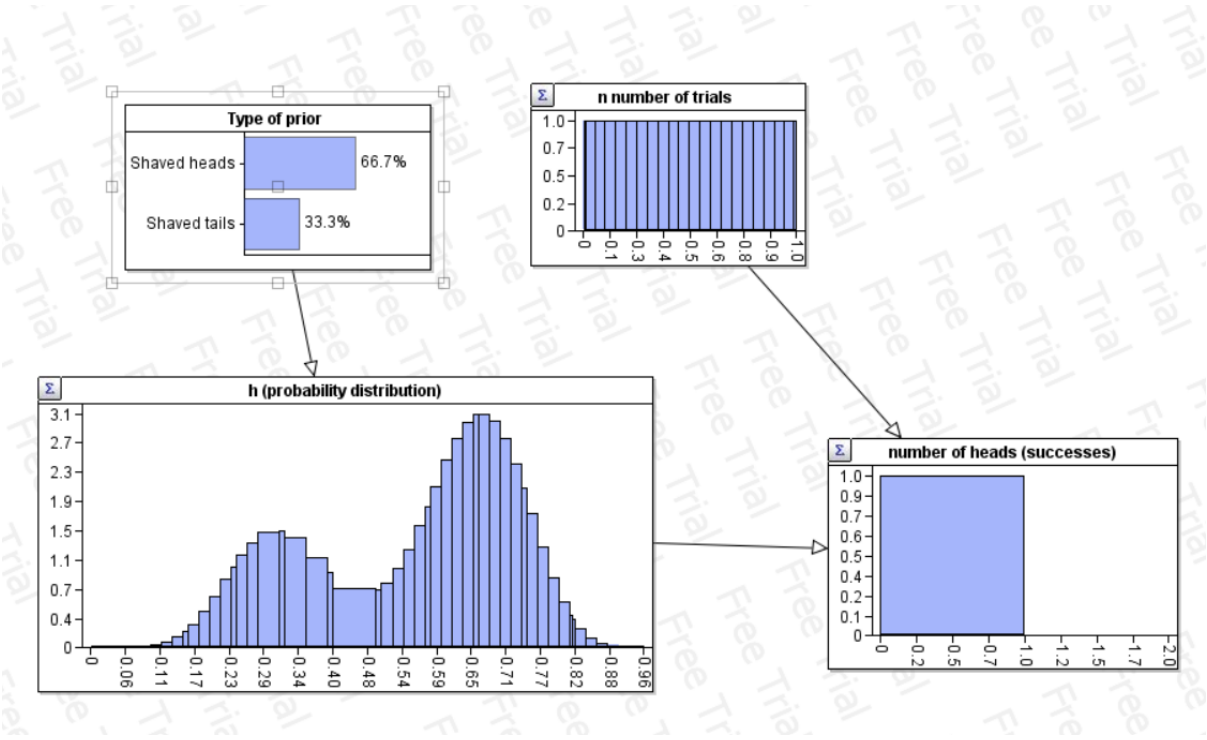

Fig. 9: AgenaRisk Model: coin biased towards heads with 2:1 Probability and no data

As we can see, the distribution is not strongly defined to the right or two the left. This is because we have not added any observations. The smaller peal reflects our statement that there is 1/3 chance of the coin being shaved on the tails. The bigger peak reflects the second statement that there is 2/3 chance that the coin is shaved on heads. In the next question, we will see how the distribution changes when we add data.

3.3 Carefully describe and justify how you used your model and calculate the posterior probability of each hypothesis after having observed 25 coin spins:

H H H T H H T H H H T T H T H H T H H H H T H T H

Now we shall add the observations above: 8 tails and 17 heads. The output of the model is shown in Fig. 10.

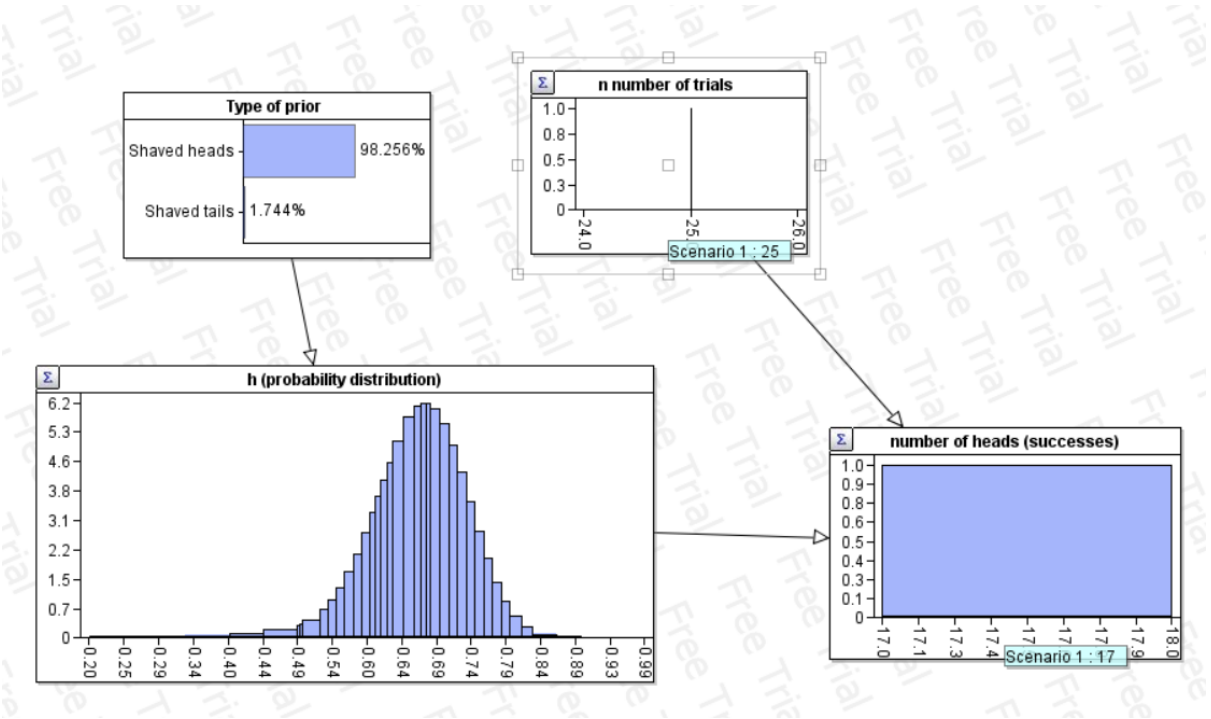

Fig. 10: AgenaRisk Model: coin biased towards heads
with 2:1 Probability and observations

As we can see, our distribution is much more centred around 0.68 and there is just one peak. In other words, it is right skewed. This means that the probability that our coin is shaved on heads is much stronger now. This is visible in the distribution graph. The distribution is much more bent and clearly right skewed.

3.4 Rather than use the binomial distribution use a series of individual single Bernoulli trials in your model and explain how your model has changed to accommodate this contrasting approach and discuss the differences, or otherwise in the result.

In the last part of the assignment, we need to create the same model but with Bernoulli nodes. Essentially, this is the same as using binomial distribution. This is because our problem is a Bernoulli problem, i.e it has only two outcomes. In other words, our Beta distribution is a distribution over the bias of a Bernoulli process, i.e shaved on heads or shaved on tails. In the end, we will get the same outcome we got in question 3.3. This is a probability of the coin being shaved on heads would be ca. 98%.

Let us now illustrate this with a model created in AgenaRisk. It is shown below in Fig.11.

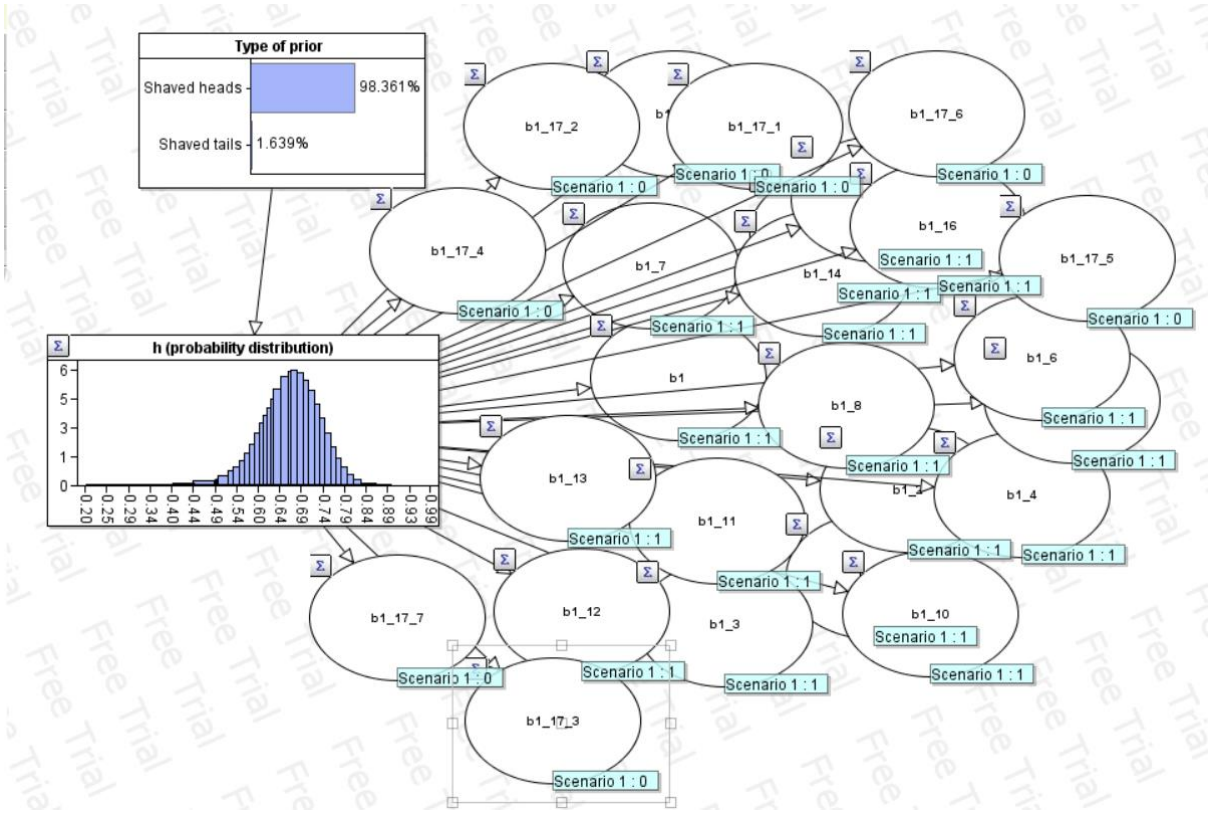

Fig. 10: AgenaRisk Model: Bernoulli approach

As we can see, we got approximately the same probability that the coin is shaved on heads.

References

1. N. Fenton, and M. Neil. 2019. Risk Assessment and Decision Analysis with Bayesian Networks: Vol. Second edition. Chapman and Hall/CRC.
2. M. Clyde, M. Çetinkaya-Rundel, C. Rundel, D. Banks, C. Chai, and L. Huang. 2021. An Introduction to Bayesian Thinking: A Companion to the Statistics with R Course. Accessed 17 March 2021, [https://statswithr.github.io/book/bayesian-inference.html#conjugacy]