Week 5

Live Discussion Session

Starts at 2.05pm

Relative versus Absolute Effectiveness: comparing 2 treatments

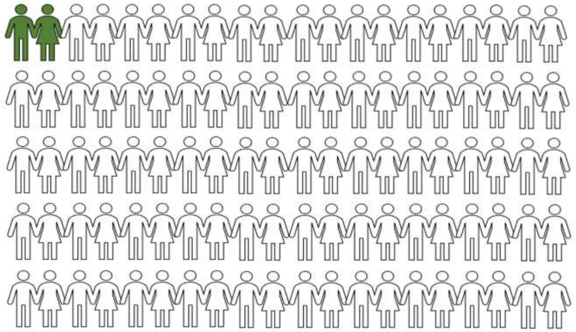
With treatment

Without treatment

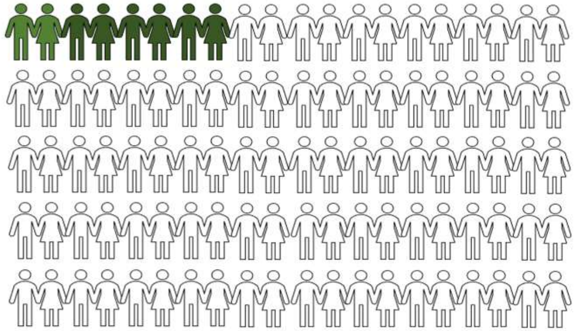
2 out of every 100 people with disease A die. Absolute risk of death is 2%

8 out of every 100 people with disease A die. Absolute risk is 8%

Treatment for disease A



Treatment A Effectiveness **75%** So 75% relative risk decrease of death with treatment

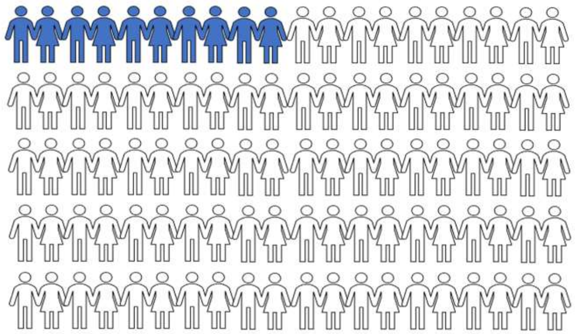


absolute risk decrease (resp. increase) of death with (resp. without) treatment is **6%**

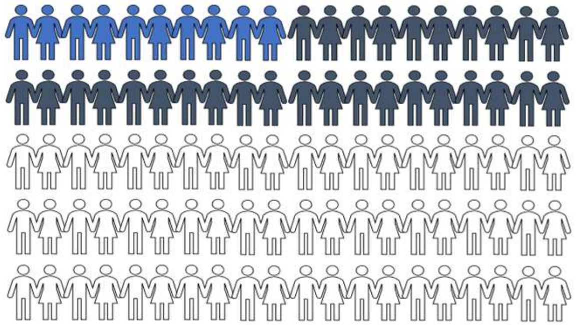
10 out of every 100 people with disease B die Absolute risk of death is 10%

40 out of every 100 people with disease B die. Absolute risk is 40%

Treatment for disease B



Treatment B Effectiveness **75%** So 75% relative risk decrease of death with treatment



absolute risk decrease (resp. increase) of death with (resp. without) treatment is **30%**

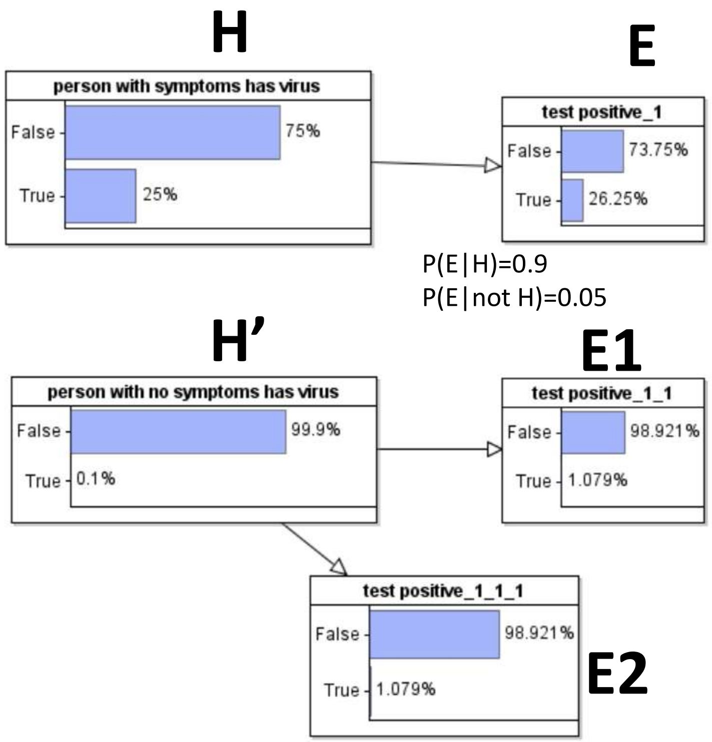
Treatments for A and B have exactly the same 'effectiveness' - but *very different absolute* risk reductions

In a choice between treating disease A or B we have to consider the population incidence rates:

	Incidence rate	Absolute risk decrease	Lives saved per 100,000	Treatment choice
Scenario 1	A: 1% B: 1%	A: 0.06% B: 0.3%	A: 600 B: 3,000	**B**
Scenario 2	A: 10% B: 1%	A: 0.6% B: 0.3%	A: 6,000 B: 3,000	**A**

Error!!!!!

Should be:
60, 300
600, 300



P(E1|H')=0.8
P(E1|not H')=0.01
(same for E2)

a) P(virus) = P(virus with symptoms)*P(symptoms) + P(virus no symptoms)*P(no symptoms)
=P(H)*P(symptoms) + P(H')*P(no symptoms)= **(0.25 x 0.1)+(0.001 x 0.9) = 0.0259 = 2.59%**
as10% of the population have symptoms

b) What is the probability a person with symptoms will test positive?

$P(E) = P(E|H) \times P(H) + P(E|not\ H) \times P(not\ H)$ =0.9 x 0.25 + 0.05 x 0.75 = 0.2625 = 26.25%

c) What is the probability a person without symptoms will test positive?

$P(E1) = P(E1|H') \times P(H') + P(E1|not\ H') \times P(not\ H')$=0.8 x 0.001+ 0.01 x 0.999 = 0.01079 = 1.079%

d) A person with symptoms tests positive. What is the probability they have the virus?

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} = \frac{0.9 \times 0.25}{0.2625} = 0.857 = 85.7\%$$

e) A person with symptoms tests negative. What is the probability they have the virus?

$$P(H|notE) = \frac{P(not\ E|H) \times P(H)}{P(not\ E)} = \frac{0.1 \times 0.25}{0.7375} = 0.0339 = 3.39\%$$

f) A person with no symptoms tests positive. What is the probability they have the virus?

$$P(H'|E1) = \frac{P(E1|H') \times P(H')}{P(E1)} = \frac{0.8 \times 0.001}{0.01079} = 0.07414 = 7.414\%$$

g) A person without symptoms tests positive. Assuming second test is independent of first, what is the probability they test positive in a second test?

We know P(H'|E1) = 0.07414 Let H" = H'|E1
Then P(E2) = P(E2|H") x P(H") + P(E2 | not H") x P(not H") = 0.8 x 0.07414 + 0.01 x 0.92596 = 0.06857 = 6.857%

h) A person without symptoms tests positive in both the first and second test. What is the probability they have the virus?
We know from f) that after the first positive test the revised posterior probability of H' is 0.07414. But from g) we know the probability of testing positive on second if the first was positive is 0.06857. So we now have revised posteriors of P(H')=0.07414 and P(E2)=0.06857

$$\boxed{P(H'|E2) = \frac{P(E2|H') \times P(H')}{P(E2)} = \frac{0.8 \times 0.07414}{0.06857} = 0.86498 = 86.498\%}$$

$$P(H'|(E1\ and\ E2)) = \frac{P((E1\ and\ E2)|H') \times P(H')}{P(E1\ and\ E2)} = \frac{P((E1\ and\ E2)|H') \times P(H')}{P((E1\ and\ E2)|H') \times P(H') + P((E1\ and\ E2)|not\ H') \times P(not\ H')}$$

$$= \frac{P(E1\ |H') \times P(E2\ |H') \times P(H')}{P(E1\ |H') \times P(E2\ |H') \times P(not\ H') + P(E1\ |not\ H') \times P(E2\ |H') \times P(H')}$$

Because E1 and E2 are independent given H’

$$= \frac{0.8 \times 0.8 \times 0.001}{0.8 \times 0.8 \times 0.001 + 0.01 \times 0.01 \times 0.999} = 0.86498 = 86.498\%$$

The general Binomial formula: it's all about knowing how to count combinations

In general we want to consider any number *n* independent "trials" (such as 5 die rolls) where on each trial the probability of "success" is *p* (e.g. *p*=1/6).

For each *r* from 0 to *n*, we want to know the probability of getting exactly *r* successes in these *n* trials.
Let's consider one specific such sequence:

$$\overbrace{S\ \ S\ \ S \dots\ S\ \ S}^{r \text{ of these}}\ \overbrace{F\ \ F\ \ F \dots\ F}^{n\text{-}r \text{ of these}}$$

The probability of getting this particular sequence of *r* successes is

$$\boldsymbol{p^r(1-p)^{n-r}}$$

The number of such sequences is the number of different ways we can arrange *r* objects into *n* positions

$$\frac{\boldsymbol{n}!}{\boldsymbol{r}! \times (\boldsymbol{n}-\boldsymbol{r})!}$$

(The number of combinations of *r* objects from *n*) also written as $\binom{n}{r}$

But there are several ways to get 2 out of 5 successes

X	X	–	–	–
X	–	X	–	–
X	–	–	X	
X	–	–	–	X
–	X	X	–	–
–	X	–	X	–
–	X	–	–	X
–	X	X	–	–
–	–	X	X	–
–	–	X	–	X
–	–	–	X	X

For example, the number of different ways we can arrange **two** S's **(so *r*=2)** into **five** positions (**so *n*=5**) is

$$\frac{5!}{2! \times 3!} = \frac{5 \times 4 \times \cancel{3 \times 2 \times 1}}{(2 \times 1) \times \cancel{(3 \times 2 \times 1)}} = 10$$

Hence the Binomial Distribution

If there are *n* independent "trials" where on each trial the probability of "success" is *p* then the probability of getting exactly *r* successes is ……..

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

So, if we roll the die 8 times where 'success' is 'roll a 6' then

P(S=0) $\frac{8!}{0! \times 8!} \times \left(\frac{1}{6}\right)^{0} \left(\frac{5}{6}\right)^{8} = 0.23257$

P(S=1) $\frac{8!}{1! \times 7!} \times \left(\frac{1}{6}\right)^{1} \left(\frac{5}{6}\right)^{7} = 0.37211$

P(S=2) $\frac{8!}{2! \times 6!} \times \left(\frac{1}{6}\right)^{2} \left(\frac{5}{6}\right)^{6} = 0.26048$

P(S=3) $\frac{8!}{3! \times 5!} \times \left(\frac{1}{6}\right)^{3} \left(\frac{5}{6}\right)^{5} = 0.10419$

P(S=4) $\frac{8!}{4! \times 4!} \times \left(\frac{1}{6}\right)^{4} \left(\frac{5}{6}\right)^{4} = 0.026048$

P(S=5) $\frac{8!}{5! \times 3!} \times \left(\frac{1}{6}\right)^{5} \left(\frac{5}{6}\right)^{3} = 0.0041676$

P(S=6) $\frac{8!}{6! \times 2!} \times \left(\frac{1}{6}\right)^{6} \left(\frac{5}{6}\right)^{2} = 0.000041676$

P(S=7) $\frac{8!}{7! \times 1!} \times \left(\frac{1}{6}\right)^{7} \left(\frac{5}{6}\right)^{1} = 0.000002382$

P(S=8) $\frac{8!}{8! \times 0!} \times \left(\frac{1}{6}\right)^{8} \left(\frac{5}{6}\right)^{0} = 0.0000000595$

..but this assumes that:

$$\binom{n}{0} = \frac{n!}{n! \times 0!} = 1$$

$$\binom{n}{n} = \frac{n!}{0! \times n!} = 1$$

.. In other words it assumes:

$$0! = 1$$

'justification' why it makes sense:

$$\frac{5!}{4!} = \frac{5 \times 4!}{4!} = 5$$

$$\frac{4!}{3!} = \frac{4 \times 3!}{3!} = 4$$

$$\frac{3!}{2!} = \frac{3 \times 2!}{2!} = 3$$

$$\frac{2!}{1!} = \frac{2 \times 1!}{1!} = 2$$

$$\frac{1!}{0!} = \frac{1! \times 0!}{0!} = ?? \quad = 1$$

$$0! = 1$$

$$\binom{n}{0} = \frac{n!}{n! \times 0!} = \frac{n!}{n! \times 1} = 1$$

$$\binom{n}{n} = \frac{n!}{0! \times n!} = \frac{n!}{1 \times n!} = 1$$

Assuming 6 in 49 ball lottery: each ticket has approx. 1/(14 million) probability of winning jackpot since there are

$$\binom{49}{6} = \frac{49!}{43! \times 6!} = 13{,}983{,}816 \qquad \text{combinations of 6 numbers from 49}$$

The probability a player buys a single ticket on two consecutive weeks and wins each the jackpot each time is

$$\left(\frac{1}{13983816}\right)^2 = \frac{1}{195{,}547{,}109{,}921{,}856} \approx \frac{1}{200\ trillion}$$

But on average each player buys 2 tickets per week so in each week the probability of winning is about 1/7million
So the probability of winning in two consecutive weeks is $(1/7\text{million})^2$

But, in a 20 year period there are 1040 weeks. So each person playing has

$$\binom{1040}{2} = \frac{1040!}{1038! \times 2!} = \frac{1040 \times 1039}{2} = 540{,}280 \qquad \text{combinations of pairs of weeks when they could win}$$

By the Binomial theorem the Probability each player wins exactly 2 times in 1040 weeks is:

$$540280 \times \left(\frac{1}{7\ million}\right)^2 \times \left(1 - \frac{1}{7\ million}\right)^{1038} \approx 1.0245 \times 10^8$$

That’s just over 1 in 100 million. Still very low

However, assume there are 60 million players.
We need to calculate the probability that at least one player wins twice in a 20-year period.
That is 1 minus the probability that NO player wins twice in 20 years.
The probability a player does NOT win twice in 20 years is:

$$1 - (1.0245 \times 10^{8}) = 0.9999999889755$$

By the Binomial Theorem the probability 60 million players all fail to win twice in 20 years is

$$\boxed{0.9999999889755^{60000000} = 0.516}$$

n=60000000, r=0, p=0.99999988975

So the probability at least one player wins twice in 20 years is

$$1 - 0.516 = 0.484$$

In other words there is almost a 50% chance.
Much better then the chance of rolling a 6 on a fair die

Risk ratios, odds ratios, and hazard ratios

D = person has the disease
R = person exposed to (specific) risk

$$Risk\ ratio = \frac{P(D\,|R)}{P(D|not\ R)} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Also called the relative risk ratio
(likelihood ratio of the factor R on D)

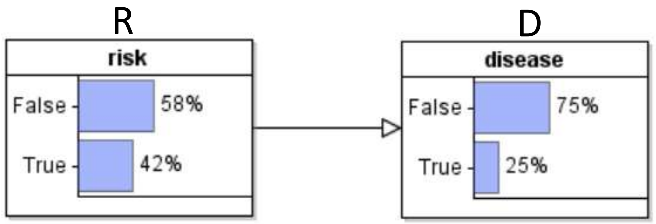
	R	Not R
D	a	c
Not D	b	d

$$Risk\ ratio = \frac{120/420}{130/580} = \frac{0.2857}{0.2241} = 1.27$$

(as it's > 1 it does indeed increase risk)

	R	Not R
D	120	130
Not D	300	450

$$relative\ risk\ (increase) = \frac{\frac{a}{a+b} - \frac{c}{c+d}}{\frac{c}{c+d}} = \frac{0.2857 - 0.2241}{0.2241} = 0.27 = risk\ ratio\ - 1 = 27\%$$



$$absolute\ risk\ (increase) = P(D\,|R) - P(D|not\ R) = \frac{a}{a+b} - \frac{c}{c+d}$$

$$absolute\ risk\ (increase) = 0.2857 - 0.2241 = 0.0616 = 6.16\%$$

$$Odds\ ratio = \frac{P(D\,|R)/P(not\ D\,|R)}{P(D|not\ R)/P(not\ D|not\ R)} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$Odds\ ratio = \frac{120/300}{130/450} = \frac{0.4}{0.29} = 1.38$$

Risk ratios and odds ratios are cumulative over an entire study, using a defined endpoint. In contrast **Hazard ratios** represent instantaneous risk over the study time period, or some subset.
For example, in a drug study, the treated population may die at twice the rate per unit time of the control population. The hazard ratio would be 2, indicating higher hazard of death from the treatment.
See: https://s4be.cochrane.org/blog/2016/04/05/tutorial-hazard-ratios/

Question 3, 2021 Exam

It is known that about 2.3% of people who have sleeping disorders have severe insomnia (defined as going more than 36 hours without being able to sleep at all)

A study of 1000 people who have sleeping disorders discovered that tea-drinkers (classified as those who drink more than 2 cups of tea a day) are more likely to suffer severe insomnia.

	Tea-drinkers	Not tea-drinkers
Severe insomnia	9	14
Other sleeping disorders	291	686
Total	300	700

a Answer the following about people with sleeping disorders:

i) What is the relative increase in risk of having severe insomnia for tea drinkers compared to non-tea drinkers? **[3 marks]**
ii) What is the absolute increase in risk of having severe insomnia for tea drinkers compared to those who are not tea-drinkers? **[3 marks]**

b Suppose we know that 10% of the population have sleep disorders. Of those with sleeping disorders, 30% are tea—drinkers. Of those with no sleeping disorders only 20% are tea drinkers. Answer the following questions about the whole population:
i) What is the relative increase in risk of having severe insomnia for tea-drinkers compared to those who are not tea-drinkers? **[5 marks]**
ii) What is the absolute increase in risk of having severe insomnia for tea drinkers compared to those who are not tea-drinkers? **[5 marks]**
Hint: *you should assume a population size of 100,000 and create two tables like above for people with and without sleep disorders.*

c What paradox could be triggered if you used the above 1000-person study to make inferences about the risk of severe insomnia caused tea-drinking to the entire population? **[2 marks]**

d Which of the following headlines is the ***most*** misleading? **[2 marks]**

i) “Study shows people with sleeping disorders should consider cutting down on the amount of tea they drink”.
ii) “Drinking more than 2 cups of tea a day leads to 50% increase in risk of severe sleep disorder”.
iii) “People with sleeping disorders who drink more than 2 cups of tea a day are at increased risk of the most severe sleep deprivation”.
iv) “Drinking more than 2 cups of tea a day may lead to severe sleep deprivation”.

a) People in the study

i) 9 out of 300 (3%) of tea drinkers have severe insomnia, 14 out of 700 (2%) non-tea drinkers 2%, so 50% relative risk increase ((3-2)/2)

ii) Absolute risk increase is 1% (3-2)

	Tea-drinkers	Not tea-drinkers
Severe insomnia	9	14
Other sleeping disorders	291	686
Total	300	700

b) Whole population (100,000): we know 10% have sleep disorders. Of those with sleeping disorders, 30% are tea—drinkers. Of those with no sleeping disorders only 20% are tea drinkers.

	Sleep disorders (10,000)		No Sleep disorders (90,000)	
	Tea drinkers (3,000)	Non-tea drinkers (7,000)	Tea drinkers (18,000)	Non-tea drinkers (72,000)
Severe	90	140	0	0
Not severe	2910	6,860	18,000	72,000

i) Relative risk increase: 90 out of 21,000 tea drinkers (=0.4286%) have severe insomnia; 140 out of 79,000 non-tea drinkers (=0.1772%) have severe insomnia. So relative risk increase is (0.4286-0.1772)/0.1772= 142%

ii) But absolute risk increase is just 0.2514%

	Sleep disorders (100,000)	
	Tea drinkers (21,000)	Non-tea drinkers (79,000)
Most Severe	90	140
Not severe	20910	78,860

c) Paradox: the 1000 person study is not representative of the population - Berkson's or Collider paradox (sample not representative of whole population)

d) (ii) is the ***most*** misleading?

i) "Study shows people with sleeping disorders should consider cutting down on the amount of tea they drink".

ii) "Drinking more than 2 cups of tea a day leads to 50% increase in risk of severe sleep disorder".

iii) "People with sleeping disorders who drink more than 2 cups of tea a day are at increased risk of the most severe sleep deprivation".

iv) "Drinking more than 2 cups of tea a day may lead to severe sleep deprivation".

Randomized controlled experiment: for people with disease D

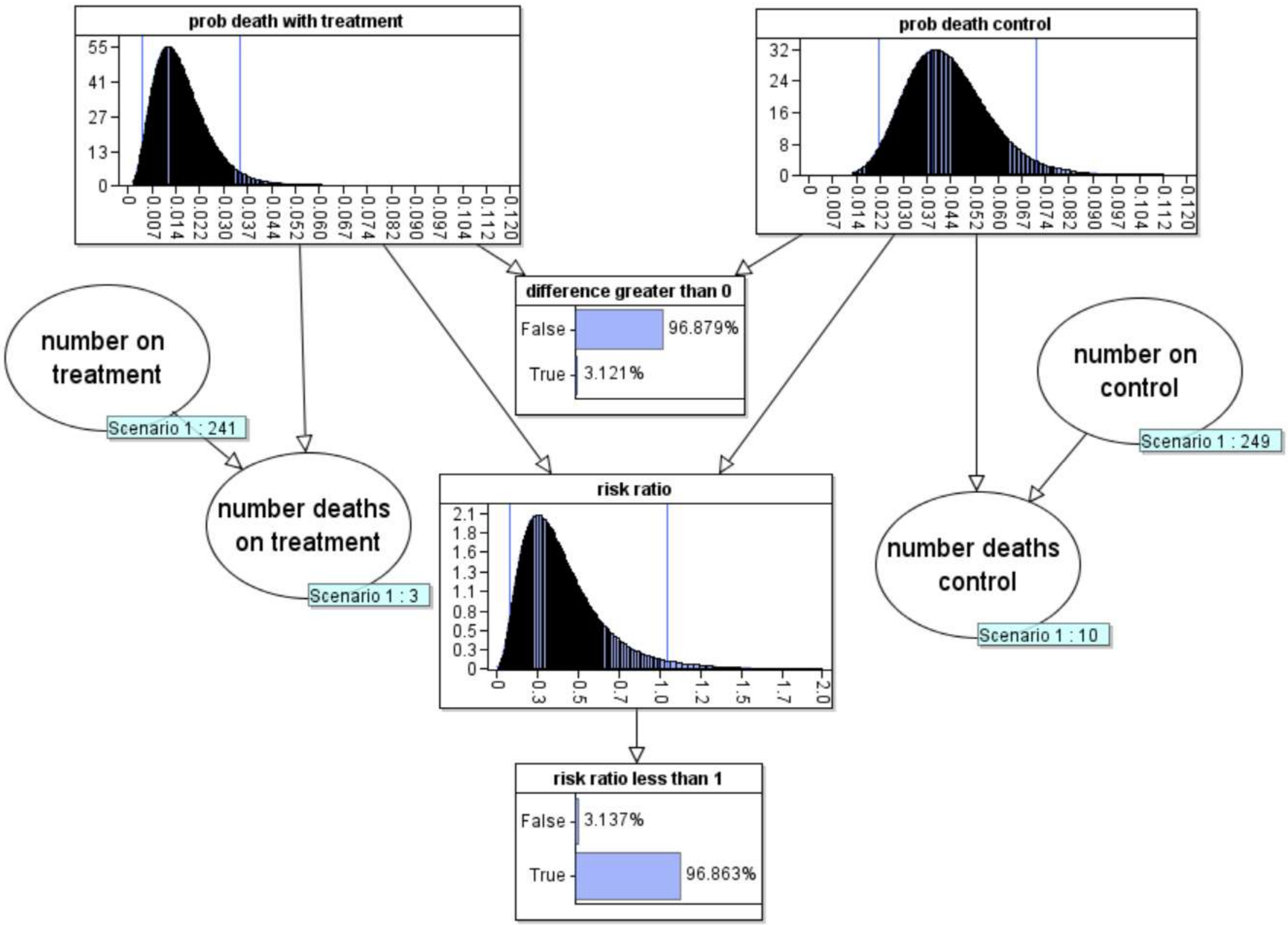
3 out of 241 who get Treatment T die
10 out of 249 who don't get Treatment T die

"Risk ratio" for treatment T on D is

$$\frac{\left(\frac{3}{241}\right)}{\left(\frac{10}{249}\right)} = \frac{0.0124}{0.0402} = 0.308$$

So risk ratio is less than 1 meaning treatment reduces mortality risk

But is it 'significant'


prob death with treatment
prob death control
difference greater than 0
False 96.879%
True 3.121%
number on treatment
Scenario 1 : 241
number deaths on treatment
Scenario 1 : 3
risk ratio
number on control
Scenario 1 : 249
number deaths control
Scenario 1 : 10
risk ratio less than 1
False 3.137%
True 96.863%

Bayes v classical stats

One double headed coin in bag with 4 fair coins

Coin is tossed twice. Both Heads. What is the probability the coin is fair?

Classical stats (two extreme positions, neither makes sense):

- 2 heads out of 2 tosses. Probability of heads = 1. Hence coin 'must be double headed"; or
- Not enough data to reach any 'statistically significant' conclusion

But of course we must revise our belief in P(fair coin) after observing even this small amount of evidence

Let H be "fair coin", E = 2 heads out of 2
P(H) = 4/5   P(E|H) = ¼   P(E| not H) = 1

$$PH|E) = \frac{P(E|H) \times P(H)}{P(E|H) \times P(H) + P(E|not\ H) \times P(not\ H)} = \frac{\frac{1}{4} \times \frac{4}{5}}{\frac{1}{4} \times \frac{4}{5} + \frac{1}{5}} = \frac{1}{2}$$

To achieve statistical significance at p-value 0.01 (1%) we must have P(E|H)<0.01

If we observe 6 out of 6 heads P(E|H) = 1/64 = 0.015625   NOT SIGNIFICANT!!!!!

Need to observe at least 7 out of 7 heads P(E|H) = 1/128 = 0.0078125