Clinical statistics for non-statisticians: Day two

Steve Simon

Re-introduce yourself

Here’s one more interesting number about myself

- 51: I started running when I was 51.

Tell us one more interesting number about yourself.

Speaker notes

Speaker notes

I had adopted a young boy from Russia a few years earlier and I could not keep up with him. A game of tag would leave me winded after only a few minutes. So I started running every other day. I was not very fast, but I built up endurance, so I could run for almost an hour without needing to take a break.

Lately, I have been a bit too busy, but I still try to run in organized races. Mostly 5 kilometer races, but I have done a few 10 kilometer races. I am not quite up to running a half marathon yet, but it is on my list of things to do before I die.

Outline of the three day course

- Day one: Numerical summaries and data visualization
- Day two: Hypothesis testing and sampling
- Day three: Statistical tests to compare treatment to a control and regression models

My goal: help you to become a better consumer of statistics

Speaker notes

Speaker notes

Just a reminder of where you are. Day one, you learned about numerical summaries and data visualization. Today, you will see information about hypothesis testing and sampling.

Day two topics

- Hypothesis testing
  - What does a p-value tell you
  - Why you might prefer a confidence interval
  - What sample size do you need
  - How does a Bayesian data analysis differ
  - What should you do if you do not have a hypothesis to test

Speaker notes

Speaker notes

Here are more details about what you will see on hypothesis testing…

Day two topics (continued)

- Sampling
  - What do you gain with a random sample
  - When might you prefer a non-random sample
  - When should you use randomization or blinding
  - What are the benefits of matching
  - How can you ensure that your sampling approach is ethical

Speaker notes

Speaker notes

…and what you will see on sampling.

Bad quiz question

A research paper computes a p-value of 0.45. How would you interpret this p-value?

- Strong evidence for the null hypothesis
- Strong evidence for the alternative hypothesis
- Little or no evidence for the null hypothesis
- Little or no evidence for the alternative hypothesis
- More than one answer above is correct.
- I do not know the answer.

A bad confidence interval

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. What does this confidence interval tell you.

Bayesian question

A Bayesian data analysis can incorporate subjective opinions through the use of Bayes rule.

- data shrinkage.
- a prior distribution.
- a posterior distribution.
- p-values.
- I do not know the answer.

Sampling question

P-values

- Most commonly reported statistic
  - Also sharply criticized
  - Requires a research hypothesis
- Two alternatives
  - Confidence intervals

Speaker notes

Speaker notes

P-values are a fundamental tools used in most research papers, but they are coming under increasing attack in the research community. P-values are an inferential tool and require a research hypothesis. data analyses. They are possibly the most commonly reported statistics in the medical literature. Two alternatives are confidence intervals and Bayesian data analysis.

Much research is not inferential and does not have a formal research hypothesis. It is a mistake to force these studies into a hypothesis testing framework. I will cover what you should do when you do not have a formal research hypothesis.

First, you need to remember some basic definitions from your Statistics 101 class.

What is a population?

- Population: the group you wish to generalize your research results to.
- Usually people, defined in terms of
  - Demography,
  - Geography,
  - Occupation,
  - Time,
  - Care requirements

Speaker notes

Speaker notes

A population is a group that you have an interest in. You want to get a better understanding of this group, so you conduct a research study and wish to generalize the results of that study to the population.

In clinical research, a population is almost always a group of people. There are a few exceptions. Sometimes you want to characterize inanimate objects, such as a group of hospitals or a group of medical devices. But let’s keep the focus on people for now.

A population of people is defined in terms of certain characteristics. Usually it is a combination of these characteristics.

Example of a population

An example of a population would be all infants born in the state of Missouri during the 1995 calendar year who have one or more visits to the Emergency room during their first year of life.

Speaker notes

Speaker notes

Here is an example of a population. It has many of the characteristics described on the previous slide: demography (infants), geography (born in Missouri), time (born in calendar year 1995, during first year of life) and care requirements (one or more ER visits).

Most times the population is so large that it is difficult to get data on all the individuals of that population.

Here, we actually did have access to the data on all 29,637 infants, but most times you would not be so fortunate.

What is a sample?

- Sample: subset of a population.
- Random sample: every person has the same probability of being in the sample.
- Biased sample: Some people have a decreased probability of being in the sample.
  - Always ask “who was left out?”

::: notes Speaker notes

A sample is a subset of a population. Because that population of infants was so large, you decided to collect data on a smaller group, a sample of 100 infants, say.

Statistics, according to one definition is the use of data from samples to make inferences about populations. That may be a bit too narrow a definition, but it does characterize quite a bit of what we statisticians do.

A random sample is a special type of sample. It is chosen in a way to insure that every person in the sample has the same probability of being in the sample.

In contrast a biased sample is one where some people in the population have a decreased chance of being in the sample. Often in a biased sample some people in the population are totally excluded. :::

An example of a biased sample

A research wants to characterize illicit drug use in teenagers. She distributes a questionnaire to students attending a local public high school (in the U.S. high school is grades 9-12, which is mostly students from ages 14 to 18.)

Explain how this sample is biased. Who has a decreased or even zero probability of being selected.

Type your ideas in the chat box.

Speaker notes

Speaker notes.

Here is a scenario where a researcher selects a biased sample. I should note here that this is an example specific to the United States. In Italy, you might talk about a survey distributed to the scuola secondaria di secondo grado.

[Stop and get student responses.]

What is a Type I error?

- A Type I error is rejecting the null hypothesis when the null hypothesis is true (false positive).
  - Example involving drug approval: a Type I error is allowing an ineffective drug onto the market.

What is a Type II error?

- A Type II error is accepting the null hypothesis when the null hypothesis is false.
  - An example involving drug approval: a Type II error is keeping an effective drug off of the market.

What is a p-value?

A p-value is a measure of how much evidence we have against the null hypothesis.

The null hypothesis, traditionally represented by the symbol H0, represents the hypothesis of no change or no effect.

The smaller the p-value, the more evidence we have against H0.

The p-value is also a measure of how likely we are to get a certain sample result or a result “more extreme,” assuming H0 is true.

The type of hypothesis (right tailed, left tailed or two tailed) will determine what “more extreme” means.

What is a confidence interval?

A confidence interval is a range of values that tries to quantify uncertainty associated with the sampling process.

Consider it as a range of plausible values.

There is a confidence level associated with any confidence interval, usually 95%, but sometimes 90% or 99%.

The confidence level is related to the alpha level (probability of a Type I error).

It also has a long range sampling interpretation.

If you repeatedly sampled from the same population, then 95% (or 90% or 99%) of the confidence intervals produced would contain the true value in the population.

Bayesian example

There's a wonderful example of Bayesian data analysis at work that is simple and fun. It's taken directly from an article by Jim Albert in the Journal of Statistics Education (1995, vol. 3 no. 3) which is available on the web at

www.amstat.org/publications/jse/v3n3/albert.html.

I want to use his second example, involving a comparison of ECMO to conventional therapy in the treatment of babies with severe respiratory failure.

In this study, 28 of 29 babies assigned to ECMO survived and 6 of 10 babies assigned to conventional therapy survived.

Refer to the Albert article for the source of the original data.

Bayes rule

Wikipedia gives a nice general introduction to the concept of Bayesian data analysis with the following formula:

$P(H|E) = P(E|H)\ P(H)\ /\ P(E)$

where H represents a particular hypothesis, and E represents evidence (data). P, of course, stands for probability.

Prior

Speaker notes

The first step is to specify P(H), which is called the prior probability. Specifying the prior probability is probably the one aspect of Bayesian data analysis that causes the most controversy. The prior probability represents the degree of belief that you have in a particular hypothesis prior to collection of your data.

Subjective prior

Speaker notes

The prior distribution can incorporate data from previous related studies or it can incorporate subjective impressions of the researcher. What!?! you're saying right now. Aren't statistics supposed to remove the need for subjective opinions?

Actually, a bit of subjectivity is a good thing.

First, it is impossible to totally remove subjective opinion from a data analysis. You can't do research without adopting some informal rules. These rules may be reasonable, they may be supported to some extent by empirical data, but they are still applied in a largely subjective fashion.

Here are some of the subjective beliefs that I use in my work:

- you should always prefer a simple model to a complex model if both predict the data with the same level of precision.
- you should be cautious about any subgroup finding that was not pre-specified in the research protocol.
- if you can find a plausible biological mechanism, that adds credibility to your results.

Second, the use of a range of prior distributions can help resolve controversies involving conflicting beliefs. For example, an important research question is whether a research finding should "close the book" to further research. If data indicates a negative result, and this result is negative even using an optimistic prior probability, then all researchers, even those with the most optimistic hopes for the therapy, should move on.

Third, while Bayesian data analysis allows you to incorporate subjective opinions into your prior probability, it does not require you to incorporate subjectivity. Many Bayesian data analyses use what it called a diffuse or non-informative prior distribution. This is a prior distribution that is neither optimistic nor pessimistic, but spreads the probability more or less evenly across all hypotheses.

ECMO prior

Speaker notes

Here's a simple example of a diffuse prior that Dr. Albert used for the ECMO versus conventional therapy example. Let's assume that the true survival rate could be either 0, 10%, 20%, …, 100% in the ECMO group and similarly for the conventional therapy group. This is not an optimal assumption, but it isn't terrible either, and it allows us to see some of the calculations in action.

With 11 probabilities for ECMO and 11 probabilities for conventional therapy, we have 121 possible combinations. How should we arrange those probabilities? One possibility is to assign half of the total probability to combinations where the probabilities are the same for ECMO and conventional therapy and the remaining half to combinations where the probabilities are different. Split each of these probabilities evenly over all the combinations.

This is P(H). For simplicity, we multiplied each probability by 1000 and rounded.

Layout the ECMO and control probabilities

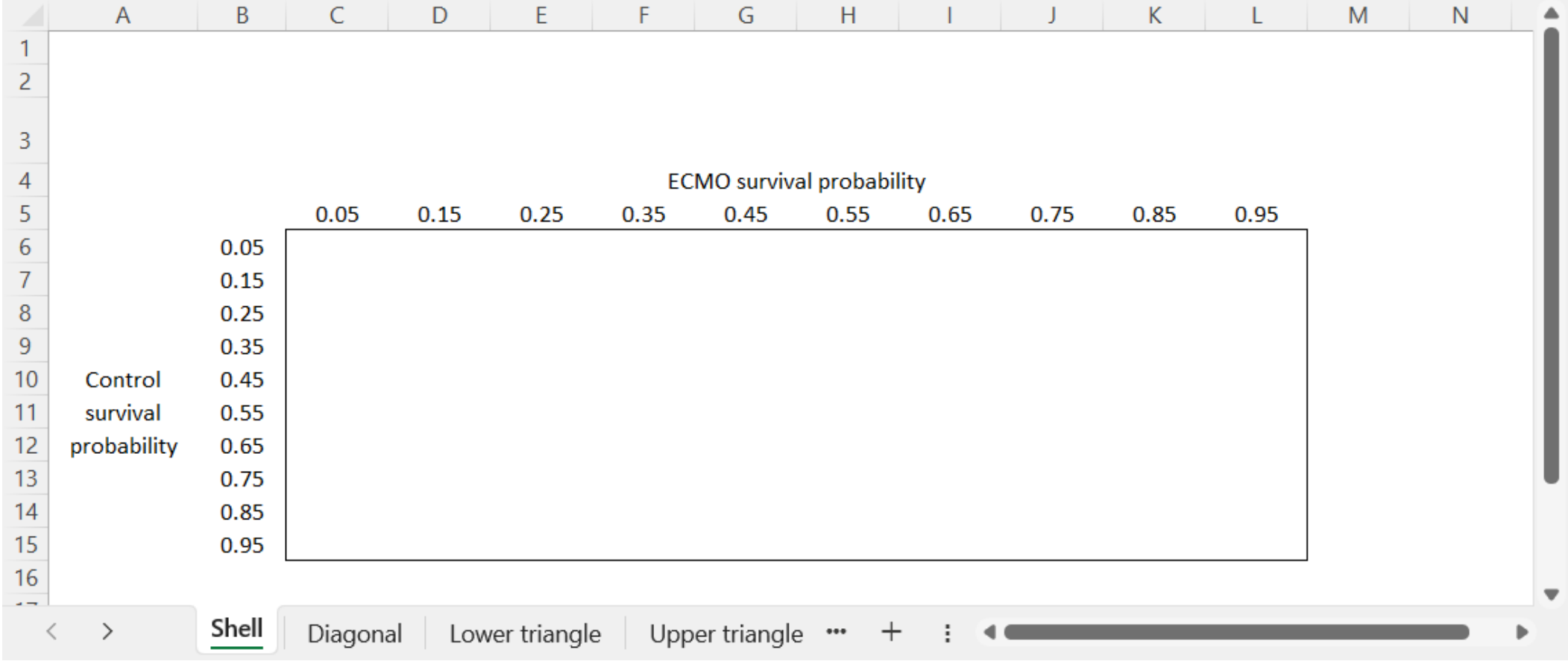


Figure 1: Empty table for prior probabilities

Place part of the probability on the diagonal

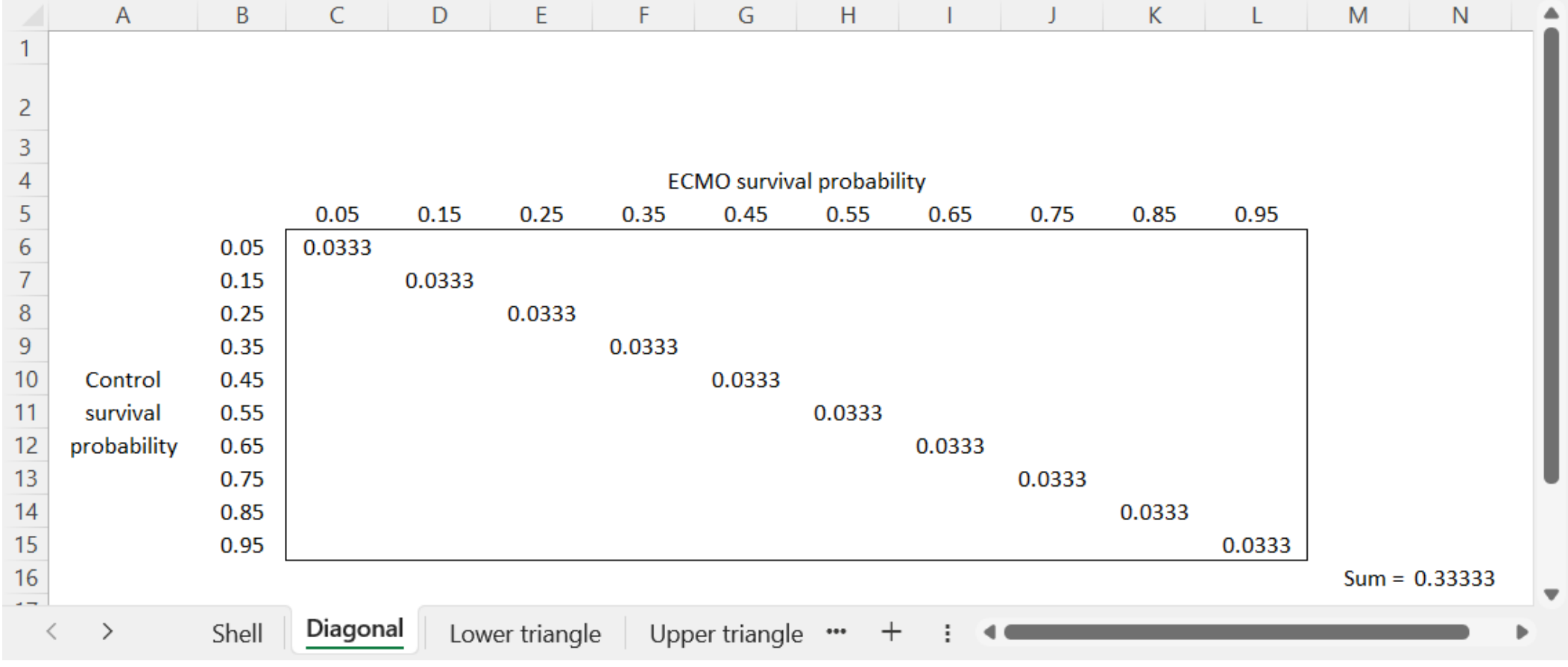
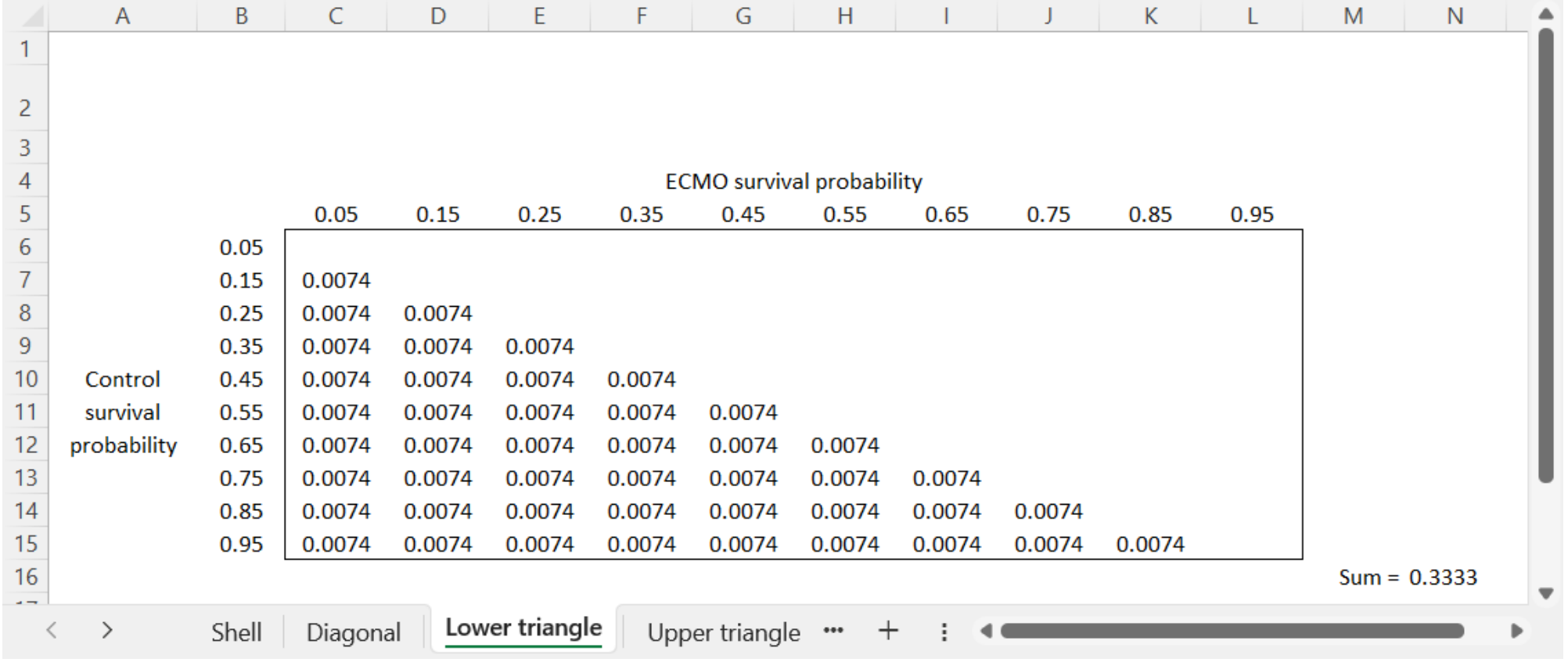


Figure 2: Table with diagonal priors

Place part of the probabilty below the diagonal

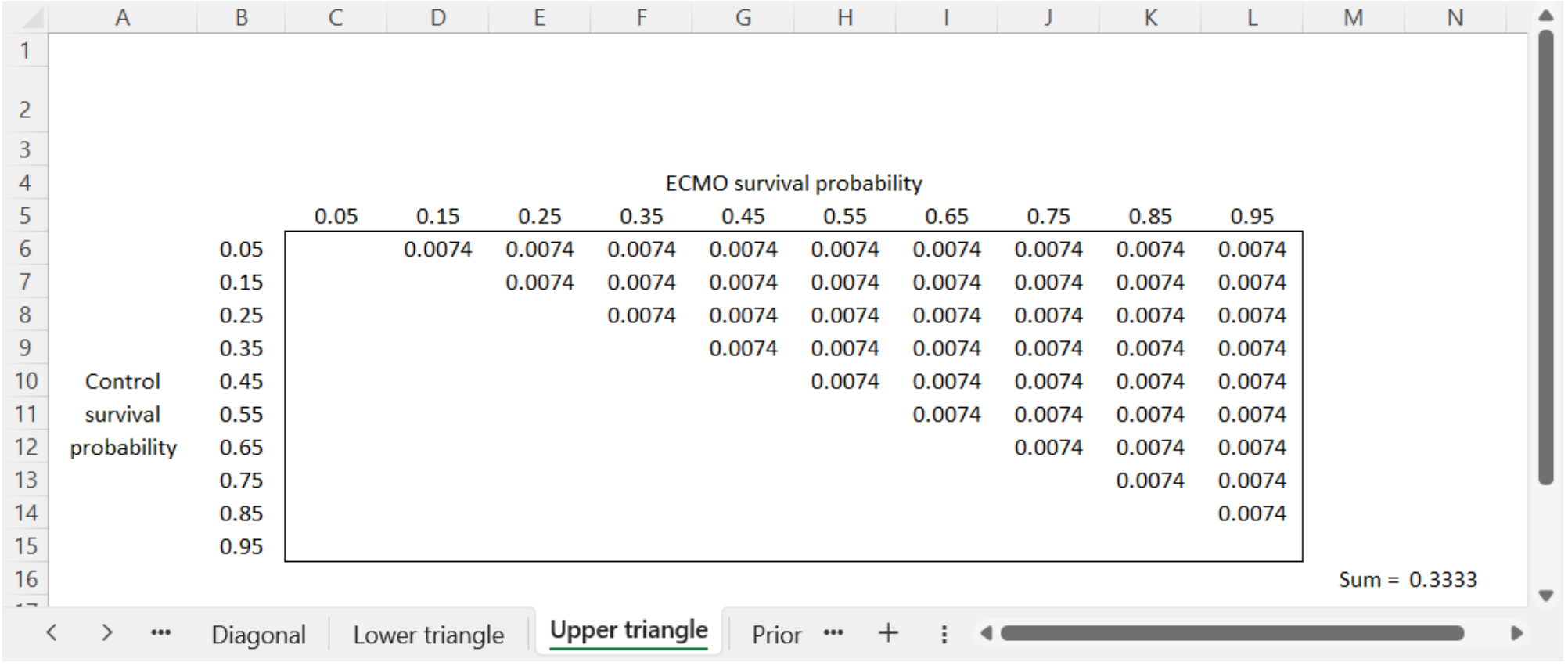


		ECMO survival probability									
		0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
	0.05										
	0.15	0.0074									
	0.25	0.0074	0.0074								
	0.35	0.0074	0.0074	0.0074							
Control	0.45	0.0074	0.0074	0.0074	0.0074						
survival	0.55	0.0074	0.0074	0.0074	0.0074	0.0074					
probability	0.65	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074				
	0.75	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074			
	0.85	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074		
	0.95	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	

Sum = 0.3333

Shell | Diagonal | Lower triangle | Upper triangle

Figure 3: Table with lower triangle of prior probabilities

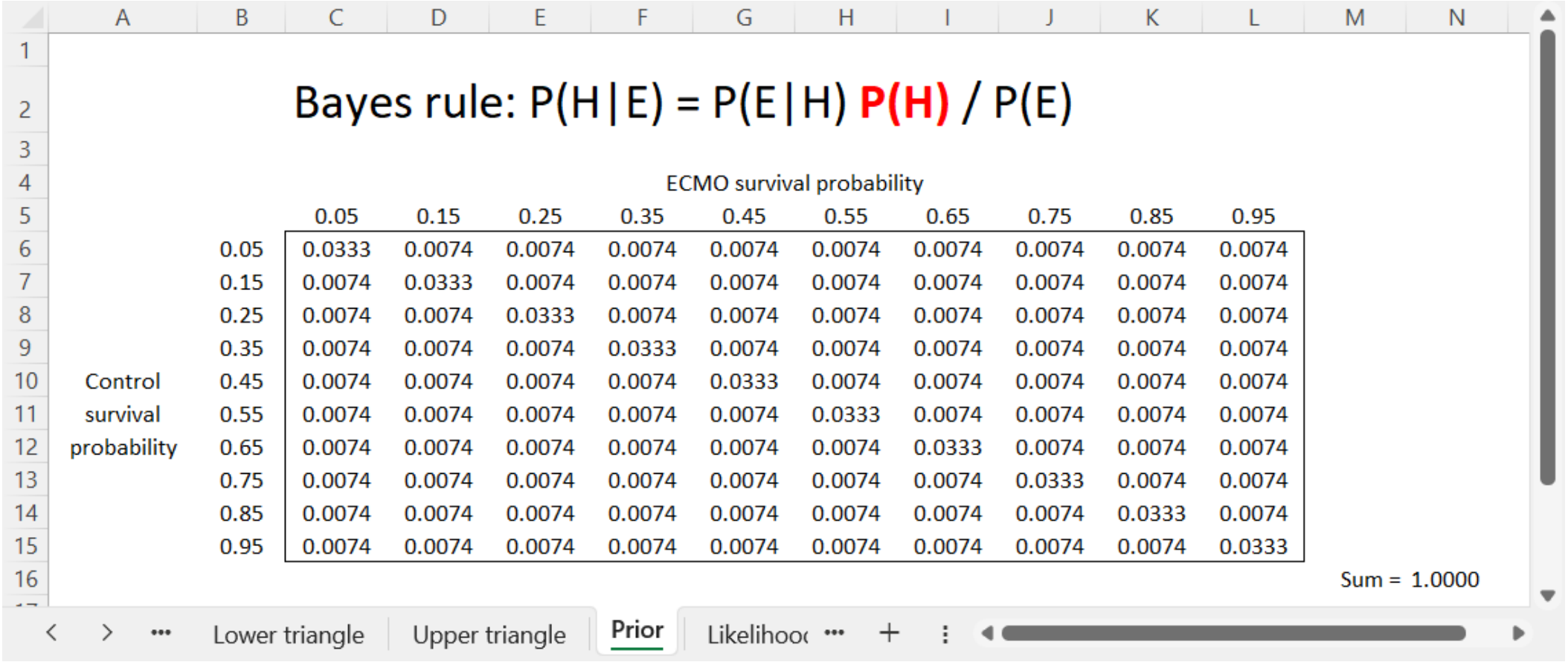


		ECMO survival probability								
		0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Control survival probability	0.05		0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074
	0.15			0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074
	0.25				0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074
	0.35					0.0074	0.0074	0.0074	0.0074	0.0074	0.0074
	0.45						0.0074	0.0074	0.0074	0.0074	0.0074
	0.55							0.0074	0.0074	0.0074	0.0074
	0.65								0.0074	0.0074	0.0074
	0.75									0.0074	0.0074
	0.85										0.0074
	0.95										

Sum = 0.3333

Figure 4: Table with upper triangle of prior probabilities

Prior table



Bayes rule: $P(H|E) = P(E|H)\,\mathbf{P(H)} / P(E)$

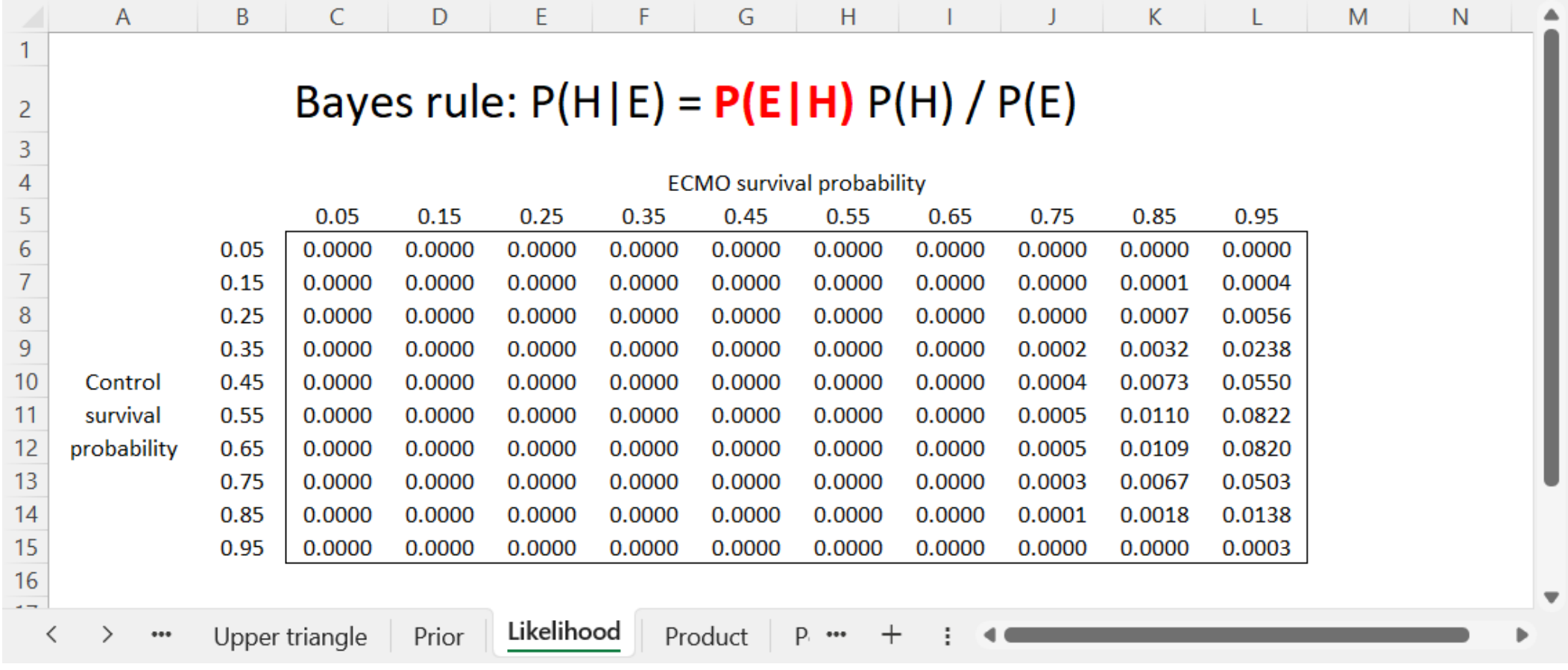
Control survival probability	ECMO survival probability 0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
0.05	0.0333	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074
0.15	0.0074	0.0333	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074
0.25	0.0074	0.0074	0.0333	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074
0.35	0.0074	0.0074	0.0074	0.0333	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074
0.45	0.0074	0.0074	0.0074	0.0074	0.0333	0.0074	0.0074	0.0074	0.0074	0.0074
0.55	0.0074	0.0074	0.0074	0.0074	0.0074	0.0333	0.0074	0.0074	0.0074	0.0074
0.65	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0333	0.0074	0.0074	0.0074
0.75	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0333	0.0074	0.0074
0.85	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0333	0.0074
0.95	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0074	0.0333

Sum = 1.0000

Lower triangle | Upper triangle | Prior | Likelihoo

Figure 5: Complete table of prior probabilities

Calculate likelihood for each ECMO/control probability



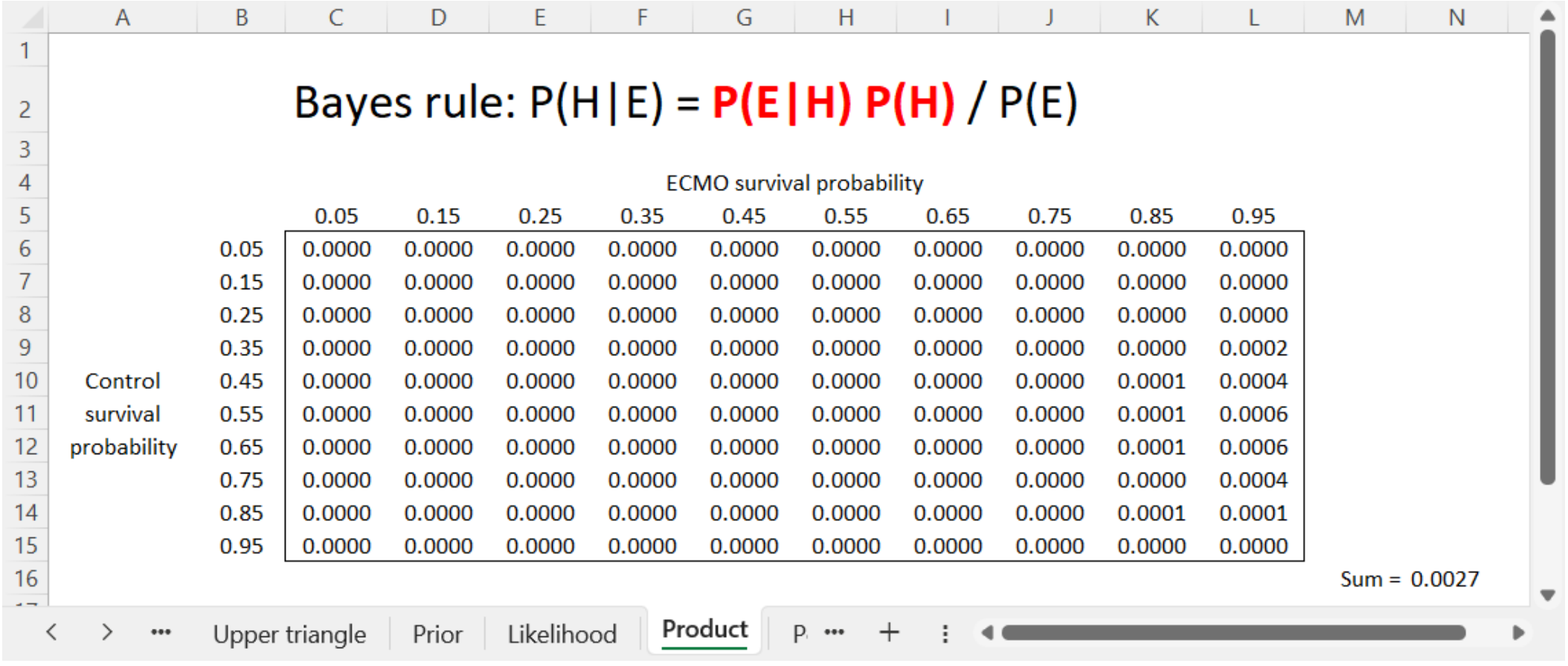
Bayes rule: $P(H|E) = \mathbf{P(E|H)}\ P(H) / P(E)$

		ECMO survival probability									
		0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Control survival probability	0.05	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004
	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0056
	0.35	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0032	0.0238
	0.45	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0073	0.0550
	0.55	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0110	0.0822
	0.65	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0109	0.0820
	0.75	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0067	0.0503
	0.85	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0018	0.0138
	0.95	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003

Upper triangle | Prior | Likelihood | Product | P

Figure 6: Table of likelihoods

Multiply prior and likelihood



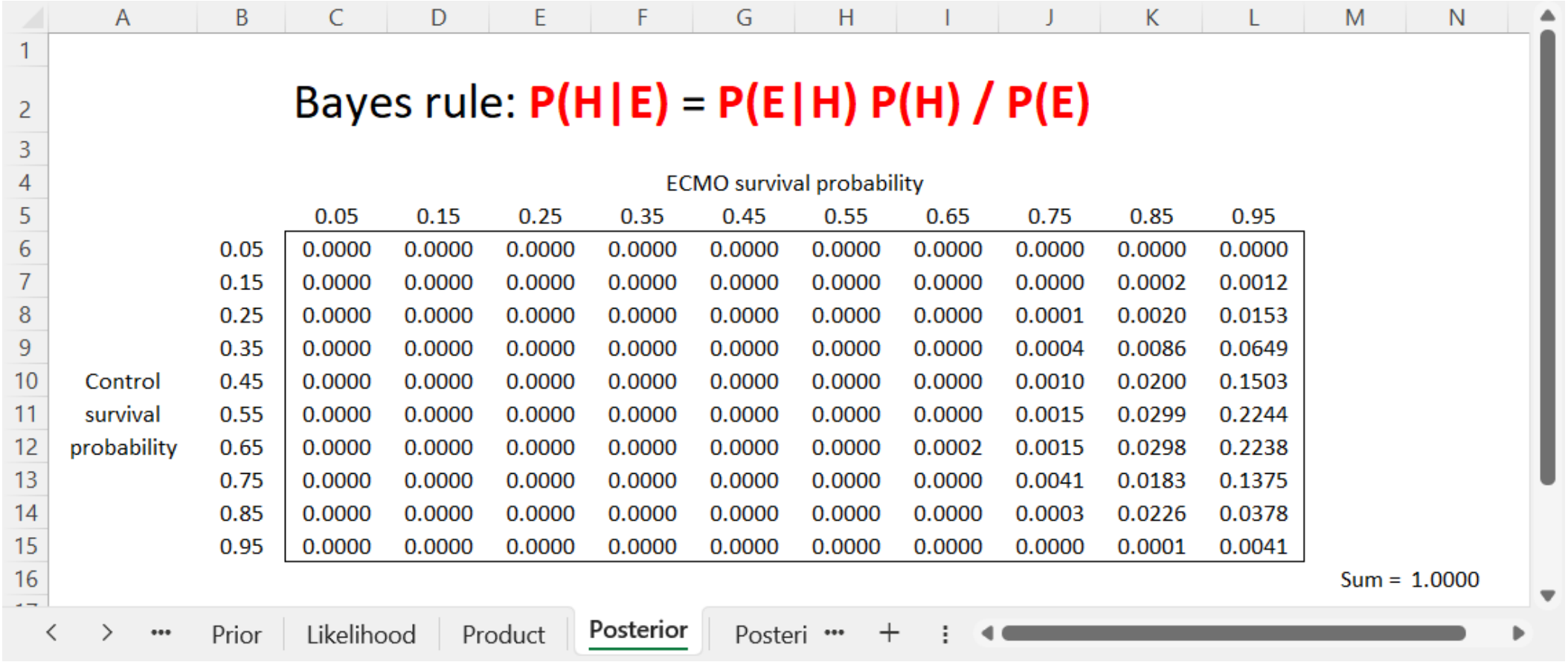
Bayes rule: $P(H|E) = P(E|H)\,P(H) / P(E)$

		ECMO survival probability									
		0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Control survival probability	0.05	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.35	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
	0.45	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004
	0.55	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006
	0.65	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006
	0.75	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004
	0.85	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
	0.95	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Sum = 0.0027

Figure 7: Product of prior and likelihood

Divide by sum to get posterior



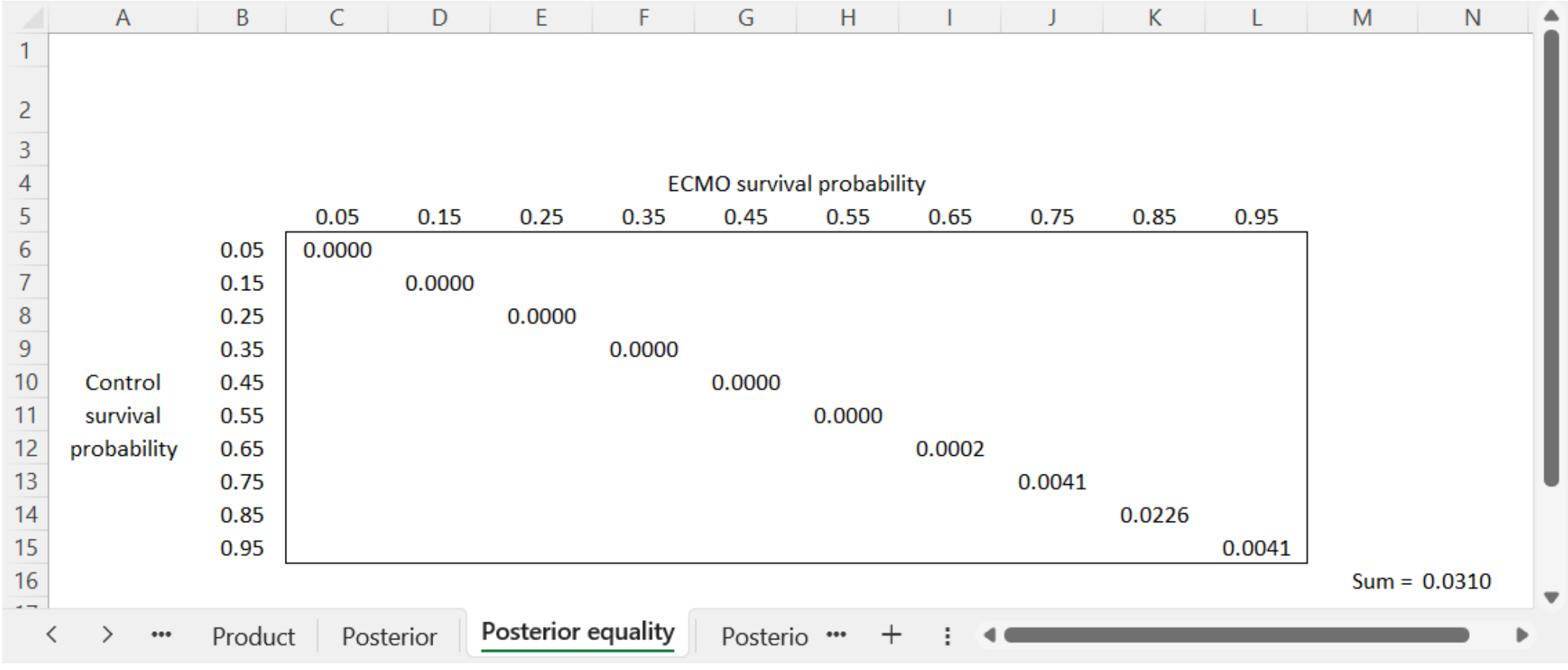
Bayes rule: $P(H|E) = P(E|H)\,P(H) / P(E)$

		ECMO survival probability									
		0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Control survival probability	0.05	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012
	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0020	0.0153
	0.35	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0086	0.0649
	0.45	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010	0.0200	0.1503
	0.55	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0015	0.0299	0.2244
	0.65	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0015	0.0298	0.2238
	0.75	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0041	0.0183	0.1375
	0.85	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0226	0.0378
	0.95	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0041

Sum = 1.0000

Figure 8: Table of prior probabilities

Posterior probability of equality

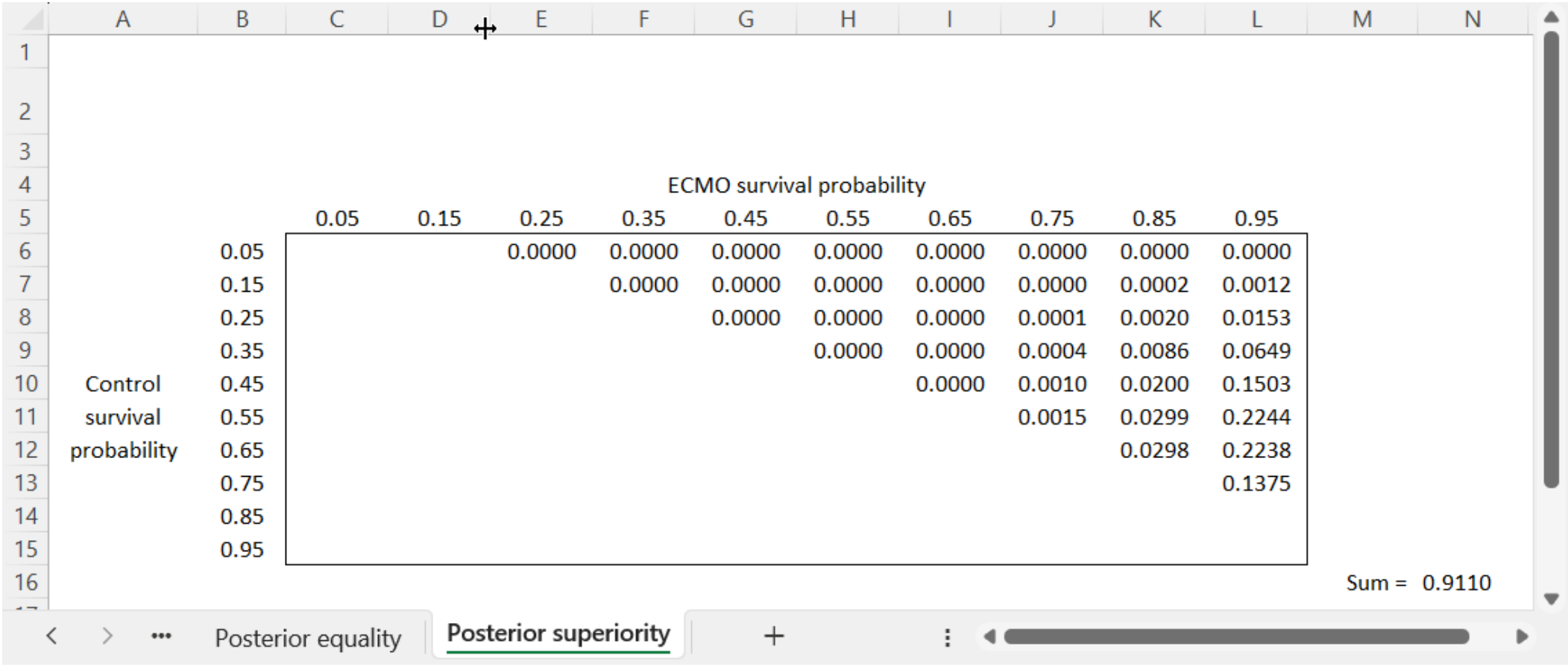


		ECMO survival probability									
		0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Control survival probability	0.05	0.0000									
	0.15		0.0000								
	0.25			0.0000							
	0.35				0.0000						
	0.45					0.0000					
	0.55						0.0000				
	0.65							0.0002			
	0.75								0.0041		
	0.85									0.0226	
	0.95										0.0041

Sum = 0.0310

Figure 9: Table of prior probabilities

Posterior probability of 20% or greater superiority



		ECMO survival probability								
		0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Control survival probability	0.05			0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.15				0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012
	0.25					0.0000	0.0000	0.0000	0.0001	0.0020	0.0153
	0.35						0.0000	0.0000	0.0004	0.0086	0.0649
	0.45							0.0000	0.0010	0.0200	0.1503
	0.55								0.0015	0.0299	0.2244
	0.65									0.0298	0.2238
	0.75										0.1375
	0.85										
	0.95										

Sum = 0.9110

Posterior equality | Posterior superiority

Table of superiority posterior probabilities

What do you gain with a random sample

When might you prefer a non-random sample

When should you use randomization or blinding

What are the benefits of matching

How can you ensure that your sampling approach is ethical

Repeat quiz questions