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

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

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

Day two topics (continued)

- You may see this on day 3 instead.
- 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

…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?

1. Strong evidence for the null hypothesis
2. Strong evidence for the alternative hypothesis
3. Little or no evidence for the null hypothesis
4. Little or no evidence for the alternative hypothesis
5. More than one answer above is correct.
6. I do not know the answer.

Speaker notes

Speaker notes

Here’s a quiz question that I proposed in an earlier presentation on p-values and confidence intervals. I wrote the responses without thinking, but then realized

“This is an easy mistake to make.”

and kept using this question. None of the answers listed above are correct, and you will see why in just a little bit.

A bad confidence interval

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. This confidence interval tells that the result is

1. statistically significant and clinically important.
2. not statistically significant, but is clinically important.
3. statistically significant, but not clinically important.
4. not statistically significant, and not clinically important.
5. The result is ambiguous.
6. I do not know the answer.

Speaker notes

Speaker notes

Here’s another question. It’s actually a good question about a bad confidence interval. If you learn nothing else today, but you understand why this is a bad confidence interval, you will have learned a lot.

By the way, the calculation of this hypothetical confidence interval is fine. It is what this confidence interval says about the researcher that matters.

Bayesian question

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

1. data shrinkage.
2. a prior distribution.
3. a posterior distribution.
4. p-values.
5. I do not know the answer.

Speaker notes

Here’s a good question. Try to answer it. If you do not know the answer, it’s okay to say 5.

P-values

- Most commonly reported statistic
  - Also sharply criticized
  - Requires a research hypothesis
- Two alternatives
  - Confidence intervals
  - Bayesian analysis
- What to do when no research hypothesis

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. 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: a group that you wish to generalize your research results to. it is defined in terms of
  - Demography,
  - Geography,
  - Occupation,
  - Time,
  - Care requirements,
  - Diagnosis,
  - Or some combination of the above.

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

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?”

Speaker 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 researcher 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

There are a variety of responses here. The sample does not include home schooled students, students in private schools, students with chronic diseases that force frequent school absences, and students who have dropped out.

Fixing a biased sample

- Redfine your population
  - Not all teenagers,
    - but those attending public high schools.

What is a parameter?

- A parameter is a number computed from a sample.
  - Examples
    - Average health care cost associated with the 29,637 children
    - Proportion of these 29,637 children who died in their first year of life.
    - Correlation between gestational age and number of ER visits of these 29,637 children.
  - Designated by Greek letters (μ, π, ρ)

What is a statistic?

- A statistic is a number computed from a sample
  - Examples
    - Average health care cost associated with 100 children.
    - Proportion of these 100 children who died in their first year of life.
    - Correlation between genstational age and number of ER visits of these 100 children.
  - Designated by non-Greek letters ($\bar{X}$, $\hat{p}$, r).

What is Statistics?

- Statistics
  - The use of information from a sample (a statistic) to make inferences about a population (a parameter)
    - Often a comparison of two populations

What is the null hypothesis?

- The null hypothesis (H_0) is a statement about a parameter.
- It implies no difference, no change, or no relationship.
  - Examples
    - $H_1: \ \mu_1 - \mu_2 \neq 0$
    - $H_0: \ \pi_1 - \pi_2 \neq 0$
    - $H_0: \ \rho \neq 0$

What is the alternative hypothesis?

- The alternative hypothesis (H_1 or H_a) implies a difference, change, or relationship.
  - Examples
    - $H_1: \ \mu_1 - \mu_2 \neq 0$
    - $H_1: \ \pi_1 - \pi_2 \neq 0$
    - $H_1: \ \rho \neq 0$

Hypothesis in English instead of Greek

- Only statisticians like Greek letters
  - Translate to simple text
  - For two group comparisons
    - Safer, more effective
  - For regression models
    - Trend, association

Speaker notes

As a researcher, you should always think about your hypothesis in terms of population parameters, but your writing should use text. Translate the Greek letters to English.

If you have a hypothesis that compares two groups, look for comparative words like “safer” or “more effective”. If your hypothesis involves some type of regression model, you should consider terms like “trend” or “association”.

Use PICO

- P = patient population
- I = intervention
- C = control
- O = outcome

Example of text hypotheses (1/2)

- "… the objective of this 78-week randomised, placebo-controlled study was to determine whether treatment with nilvadipine sustained-release 8 mg, once a day, was effective and safe in slowing the rate of cognitive decline in patients with mild to moderate Alzheimer disease."
  - Lawlor B, Segurado R, Kennelly S, et al. Nilvadipine in mild to moderate Alzheimer disease: A randomised controlled trial. PLoS Med. 2018; 15(9): e1002660. DOI: 10.1371/journal.pmed.1002660

Speaker notes

Here’s an example of a two group comparison. One group gets nilvadipine and the other group gets a placebo. Safety was measured as the proportion of patients who experienced an adverse event. The researchers also measured the proportion of patients who experienced a serious adverse event. So the Greek hypothesis would involve pi’s.

Effectiveness was measured using the Alzheimer’s Disease Assessment Scale Cognitive Subscale-12 and the Clinical Dementia Rating Scale sum of boxes. Both of these outcome measurements are continuous, so the Greek hypothesis would involve mu’s.

PICO for this study

- P = patients with mild to moderate Alzheimer disease
- I = Nilvadine
- C = placebo
- O = cognitive function

Example of text hypotheses (2/2)

- "… we investigated trends in BCC incidence over a span of 20 years and the associations between incident BCC and risk factors in a total population of 140,171 participants from 2 large US-based cohort studies: women in the Nurses' Health Study (NHS; 1986–2006) and men in the Health Professionals' Follow-up Study (HPFS; 1988–2006)."
  - Wu S, Han J, Li WQ, Li T, Qureshi AA. Basal-cell carcinoma incidence and associated risk factors in U.S. women and men. Am J Epidemiol. 2013; 178(6): 890–897. DOI: 10.1093/aje/kwt073

Speaker notes

This study used a regression model, a Cox regression model, to study trends and associations, so the Greek hypotheses would involve beta’s.

PICO for this study

- P = female nurses/male health professionals
- I = various risk factors
- C = absence of various risk factors
- O = presence/absence of BCC

One-sided alternatives

- Examples
  - $H_1: \ \mu_1 - \mu_2 > 0$
  - $H_1: \ \pi_1 - \pi_2 > 0$
  - $H_1: \ \rho > 0$
- Changes in only one direction expected
- Changes in opposite direction uninteresting

Passive smoking controversy

- EPA meta-analysis of passive smoking
  - Criticized for using a one-sided hypothesis
  - Samet JM, Burke TA. Turning science into junk: the tobacco industry and passive smoking. Am J Public Health. 2001;91(11):1742–1744.

Speaker notes

Available in html format or PDF format.

Consider a study of the effects of second-hand smoke. These studies always use directional alternatives. From what we know about active cigarette smoking is that it increases the risk of cancer and cardiovascular disease. So there is no reason to expect that passive smoke exposure should be any different than active smoking. Maybe it is less toxic, because of dilution and because the smoking coming off a cigarette from one end is different than the smoke coming off the cigarette from the other end. Fair enough, but there is not reason to believe that things are so different that all of a sudden the smoke becomes protective.

Since there is no scientific basis for a protective effect of passive smoking, it makes sense to test that passive smoking has no effect versus it having an increase in bad outcomes compared to the control group. So your null hypothesis is “not harmful” and your alternative is “harmful”. The beneficial hypothesis is lumped into the null hypothesis, but no one would dare claim that passive smoking was protective.

Actually, the tobacco companies did complain that the use of a directional alternative violated the norms of science. They won in a court battle in North Carolina, but lost on appeal.

As another aside, I was involved with prayer study. We planned this study using a one-sided hypothesis (remote prayer has a positive effect on health). The Institutional Review Board suggested changing this to a two-sided hypothesis (remote prayer has either a positive or a negative effect on health). Thankfully, we did not observe an outcome in the opposite tail as that would have been very difficult to explain.

What is a decision rule? (1/3)

- Example
  - $H_0: \ \mu_1 - \mu_2 = 0$
  - $H_1: \ \mu_1 - \mu_2 \neq 0$
  - t = $(\bar{X}_1 - \bar{X}_2)$ / se
  - Accept H_0 if t is close to zero.

What is a decision rule? (2/3)

- Example
  - $H_0 : \ \pi_1 - \pi_2 = 0$
  - $H_1 : \ \pi_1 - \pi_2 \neq 0$
  - t = $(\hat{p}_1 - \hat{p}_2)$ / se
  - Accept H_0 if t is close to zero.

What is a decision rule? (3/3)

- Example
  - $H_0 : \ \rho = 0$
  - $H_1 : \ \rho \neq 0$
  - t = r / se
  - Accept H_0 if t is close to zero.

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.
- α = P[Type I error]

Speaker notes

Speaker notes

In your research, you specify a null hypothesis (typically labeled H0) and an alternative hypothesis (typically labeled Ha, or sometimes H1). By tradition, the null hypothesis corresponds to no change. When you are using Statistics to decide between these two hypothesis, you have to allow for the possibility of error. Actually, if you are using any other procedure, you should still allow for the possibility of error, but we statisticians are the only ones honest enough to admit this.

A Type I error is rejecting the null hypothesis when the null hypothesis is true.

Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type I error would be allowing an ineffective drug onto the market.

Remember that the hypotheses involve population parameters. Population parameters are impossible to compute. So you can only talk about Type I errors in an abstract sense. You will never know for certain if you have made a Type I error.

Alpha is the probability of a Type I error, and alpha is a value that you can compute. In most studies, researchers work hard to keep the probability of a Type I error low, typically at 5%.

What is a Type II error?

- A Type II error is accepting the null hypothesis when the null hypothesis is false.
  - False negative result
  - Usually computed at MCD
  - An example involving drug approval: a Type II error is keeping an effective drug off of the market.
- β = P[Type II error]
- Power = $1 - \beta$

Speaker notes

Speaker notes

A Type II error is accepting the null hypothesis when the null hypothesis is false. You should always remember that it is impossible to prove a negative. Some statisticians will emphasize this fact by using the phrase “fail to reject the null hypothesis” in place of “accept the null hypothesis.” The former phrase always strikes me as semantic overkill.

Many studies have small sample sizes that make it difficult to reject the null hypothesis, even when there is a big change in the data. In these situations, a Type II error might be a possible explanation for the negative study results.

Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type II error would be keeping an effective drug off the market.

It bears repeating that population parameters are impossible to compute. So you will never know for certain if you have made a Type I error.

Beta is the probability of a Type II error. Beta is a known quantity. Typically researchers try to keep beta small. 10% is a typical value, though in some settings, a Type II error rate as large as 20% could be tolerated.

Power is defined as 1-beta. I will talk more about power in a little bit.

What is a p-value?

- Let t =
  - $(\bar{X}_1 - \bar{X}_2)$ / se, or
  - $(\hat{p}_1 - \hat{p}_2)$ / se, or
  - r / se
- p-value = Prob of sample result, t, or a result more extreme,
  - **assuming the null hypothesis is true**
- Small p-value, reject H_0
- Large p-value, accept H_0

Speaker notes

Speaker notes

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

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.

Alternate interpretations

- Consistency between the data and the null
  - Small value, inconsistent
  - Large value, consistent
- Evidence against the null
  - Small, lots of evidence against the null
  - Large, little evidence against the null

Speaker notes

Speaker notes

There are two interpretations that I feel are more practical. You can think of the p-value as a measure of consistency between the data and the null hypothesis. A small value implies inconsistency. It is very unlikely that you will get a value like you've seen in your sample or a value more extreme under the assumption that the null hypothesis is true. So you should reject that assumption.

On the other hand if the sample results or anything more extreme has a high probability under the assumption that the null hypothesis is true, then you should feel comfortable accepting that assumption.

I have argued that the p-value is a measure of evidence. Some have called it a poor measure of evidence, but I stand by my interpretation.

If the p-value is small, you have lots of evidence against the null hypothesis. If the p-value is large, you have little or no evidence against the null hypothesis.

What the p-value is not (1/2)

- A p-value is NOT the probability that the null hypothesis is true.
  - P[t or more extreme | null] is different than
  - P[null | t or more extreme]
    - P[null] is nonsensical
    - μ, π, or ρ are unknown constants (no sampling error)

Speaker notes

Speaker notes

The p-value is a conditional probability, and you always need to be careful about conditional probabilities. It is a probability about a sample result given an assumption about the population result. It is not a probability about a population result given the sample result. There are two reasons for this.

First, you can’t reorder a conditional probability. The probability of A given B is almost never the same as the probability of B given A. The example I give for this is the probability of being happy given that you are rich. That’s a pretty high number, I hope you’ll agree. There are a few rich people who lead miserable lives, but from everything I’ve seen, most rich people are pretty darn happy. The reverse of this is the probability of being rich given that you are happy. That number is much smaller. Because although I believe that money can buy happiness, a lot of other things can also buy happiness just as well. It’s not quite as easy to find happiness if you’re poor, but somehow, a lot of poor people find a way to be happy anyway.

A second reason that you can’t reverse the order is that you cannot make a probability statement about population parameters. They are numbers computed from the entire population, and are fixed values. You cannot make a probability statement about something that has no sampling error.

Only numbers computed from a sample (i.e., statistics) have sampling error.

What the p-value is not (2/2)

- Not a measure FOR either hypothesis
  - Little evidence **against** the null $\neq$ lots of evidence **for** the null
- Not very informative if it is large
  - Need a power calculation, or
  - Narrow confidence interval
- Not very helpful for huge data sets

Speaker notes

The p-value is not a measure for either hypothesis. It is always a measure against a particular hypothesis. Now when the p-value is small, you can make a strong statement. We have lots of evidence against the null hypothesis. That translates into lots of evidence in favor of the alternative hypothesis.

When the p-value is large, however, you are in a quandary. Little or no evidence against the null hypothesis is not the same as lots of evidence for the null hypothesis.

It’s possible to have little or no evidence against the null and also have little or no evidence against the alternative. This happens whenever you have a really small sample size combined with a lot of noise.

You can’t prove a negative, so the saying goes. Well, you can prove a negative, but you have to work harder at it. A large p-value by itself is not persuasive, but if you combine it with a power calculation done prior to data collection, that’s pretty good evidence in support of the null hypothesis.

You could also combine a large p-value with a narrow confidence interval to support the null hypothesis. I’ll talk about that more in just a bit.

In general, the p-value is not very helpful for large samples. We’re seeing this more and more. Just about everything pops up as statistically significant with these huge data sets, and you can’t use the p-value to separate the important stuff from the trivial stuff. You need to look instead at the magnitude of the sample estimates and calculate how much uncertainty you can remove in your future predictions.

Pop quiz, revisited

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

1. Strong evidence **for** the null
2. Strong evidence **for** the alternative
3. Little or no evidence **for** the null
4. Little or no evidence **for** the alternative
5. More than one answer above is correct.
6. I do not know the answer.

Speaker notes

Here’s that pop quiz again. Take a look at it quickly. Note that the p-value is of evidence against the null hypothesis. So each of the first four responses is wrong.

I wrote this question quickly, so shame, shame on me. But I’ve reproduced the example because it illustrates an important point.

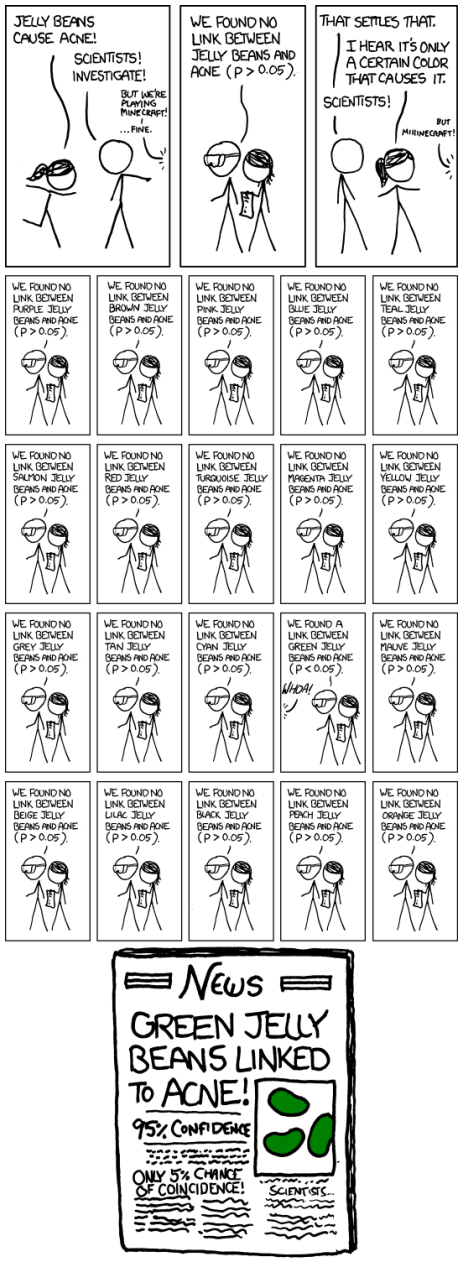


Figure 1: xkcd cartoon about jelly beans and cancer

Speaker notes

Speaker notes

This cartoon is impossible to read, but you can find it on the Canvas site or in the readings. Here's a brief run down.

In the first panel, a woman runs up to a man and shouts: Jelly beans cause acne!

The man replies : Scientists! Investigate!

In the second panel, one scientist, holding a clipboard announces: We found no link between jelly beans and acne ($p > 0.05$).

In the third panel, the woman says: I hear it's only a certain color that causes it.

In a bunch of small panels, the scientist with a clipboard reports: We found no link between purple jelly beans and acne ($p > 0.05$).

We found no link between brown jelly beans and acne ($p > 0.05$).

We found no link between pink jelly beans and acne ($p > 0.05$).

The same for blue, teal, salmon, red, and so forth. And then…

We found a link between green jelly beans and acne ($p < 0.05$). An off-screen voice goes: Whoa!

The next six panels show

We found no link between mauve jelly beans and acne ($p > 0.05$).

We found no link between beige jelly beans and acne ($p > 0.05$).

We found no link between lilac jelly beans and acne ($p > 0.05$).

We found no link between black jelly beans and acne ($p > 0.05$).

We found no link between peach jelly beans and acne ($p > 0.05$).

We found no link between orange jelly beans and acne ($p > 0.05$).

At the bottom is a newspaper with the headline: Green Jelly Beans Linked To Acne! 95% Confidence. Only 5% chance of coincidence!

If you are interested in a transcript and a detailed explanation, https://www.explainxkcd.com/wiki/index.php/882:_Significant

What is p-hacking?

- Abuse of the hypothesis testing framework.
  - Run multiple tests on the same outcome
  - Test multiple outcome measures
  - Remove outliers and retest
- Defenses against p-hacking
  - Bonferroni
  - Primary versus secondary
  - Published protocol

Speaker notes

This is an example of p-hacking. You change the testing process to increase the probability of a Type I error (Rejecting the null hypothesis when the null hypothesis is true). This increases the chance of getting a positive result, which you may find desirable, but only by increasing the probability of a false positive result.

Some examples of p-hacking. Run multiple tests on the same outcome measure. Start with the regular t-test, include the t-test that allows for unequal variances, and run two different non-parametric tests, the Wilcoxon-Mann-Whitney test and the sign test. Choose the test with the smallest p-value.

You also might consider multiple outcome measures. Compare the mortality rate, the relapse rate, and the re-hospitalization rate. If any of the three is statistically significant, claim victory.

You could also do this with longitudinal data. Compare pain relief at one hour and at four hours. If you see a difference at one hour, claim that your new medication is faster acting. If you see a difference at four hours, claim that your medication is longer lasting.

You might run a test with the full data set and then with an outlier or two removed. Report for the data set that has the smaller p-value and pretend that this was your original choice all along.

These are only a few of the choices. I don’t want to say more because I feel like I’m the devil tempting you.

There are two defenses against p-hacking. Well three if you count being honest. But what I mean is there are two things that you can do that will satisfy others that you are playing fairly.

First, you can adjust your decision rule by using a Bonferroni correction. Bonferroni divides alpha by the number of tests. If you are using three different outcome measures, compare your p-value of 0.0133 instead of 0.05.

Second, you can designate one of your outcome measures as primary. If you achieve statistical significance on your primary outcome, great. The remaining outcome measures are secondary. If you achieve statistical significance on a secondary outcome meausure only, report the results as provisional and requiring independent replication.

You should publish a detailed protocol, either through a clinical trial registry, or now there are journals which accept publications of the research protocols before any data are collected. It’s a paper with literature review and methods section, but no results and no discussion section.

Now p-hacking has happened because some people have a skewed view of research. They are interested in using research to promote their own agenda rather than using research to uncover the truth. Perfectly understandable if you are a drug company, but you as an independent researcher should never try to skew the data. It hurts you and it hurts your patients. You need to adopt a disinterested posture in that you are glad when the research points in one direction and you are glad when it points in the opposite direction, because either way, you know more than you did before and you can treat your patients better because of this knowledge.

What is a confidence interval?

- Range of plausible values
  - Tries to quantify uncertainty associated with the sampling process.

Speaker notes

Speaker notes

We statisticians have a habit of hedging our bets. We always insert qualifiers into our reports, warn about all sorts of assumptions, and never admit to anything more extreme than probable. There's a famous saying: "Statistics means never having to say you're certain."

We qualify our statements, of course, because we are always dealing with imperfect information. In particular, we are often asked to make statements about a population (a large group of subjects) using information from a sample (a small, but carefully selected subset of this population). No matter how carefully this sample is selected to be a fair and unbiased representation of the population, relying on information from a sample will always lead to some level of uncertainty.

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.

Example of a confidence interval

- Homeopathic treatment of swelling after oral surgery
  - 95% CI: -5.5 to 7.5 mm
  - Lokken P, Straumsheim PA, Tveiten D, Skjelbred P, Borchgrevink CF. Effect of homoeopathy on pain and other events after acute trauma: placebo controlled trial with bilateral oral surgery BMJ. 1995;310(6992):1439-1442.

Speaker notes

http://www.bmj.com/content/310/6992/1439.full

Always look for confidence intervals that are wide enough to drive a truck through. They are very good indicators of small sample sizes.

Consider a recent study of homoeopathic treatment of pain and swelling after oral surgery (Lokken 1995). When examining swelling 3 days after the operation, they showed that homoeopathy led to 1 mm less swelling on average. The 95% confidence interval, however, ranged from -5.5 to 7.5 mm. From what little I know about oral surgery, this appears to be a very wide interval. This interval implies that neither a large improvement due to homoeopathy nor a large decrement could be ruled out.

Now, you can’t drive a truck through an interval that goes from -5.5 to 7.5 mm, but from the perspective of a human mouth, this interval is huge. Generally when a confidence interval is very wide like this one, it is an indication of an inadequate sample size, an issue that the authors mention in the discussion section of this paper.

Confidence interval interpretation (1 of 7)

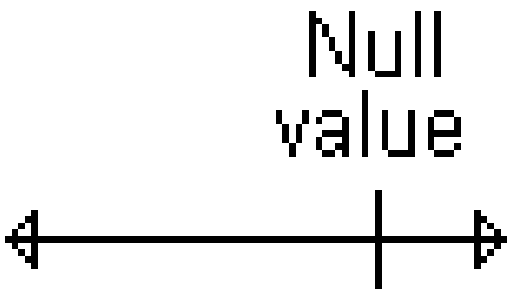


Figure 2: Interval that contains the null value

Speaker notes

When you see a confidence interval in a published medical report, you should look for two things. First, does the interval contain a value that implies no change or no effect? For example, with a confidence interval for a difference look to see whether that interval includes zero. With a confidence interval for a ratio, look to see whether that interval contains one.

Here’s an example of a confidence interval that contains the null value. This interval implies no statistically significant change.

Confidence interval interpretation (2 of 7)

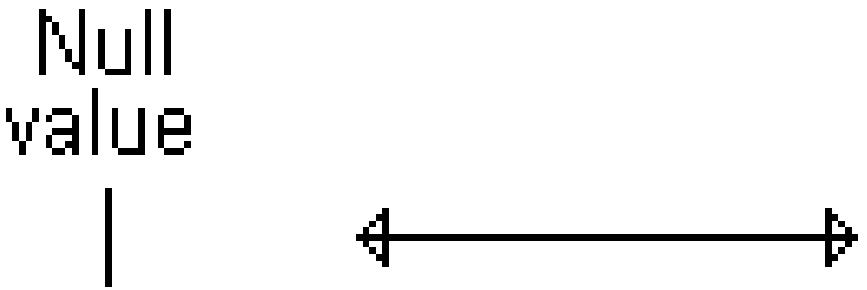


Figure 3: Interval entirely above the null value

Speaker notes

Speaker notes

Here’s an example of a confidence interval that excludes the null value. If we assume that larger implies better, then the interval would imply a statistically significant improvement.

Confidence interval interpretation (3 of 7)

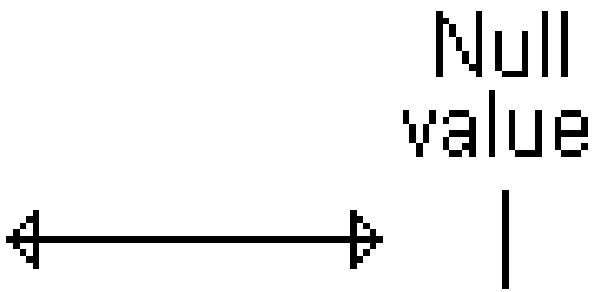


Figure 4: Interval entirely below the null value

Speaker notes

Speaker notes

Here’s a different example of a confidence interval that excludes the null value. This interval implies a statistically significant decline.

Confidence interval interpretation (4 of 7)

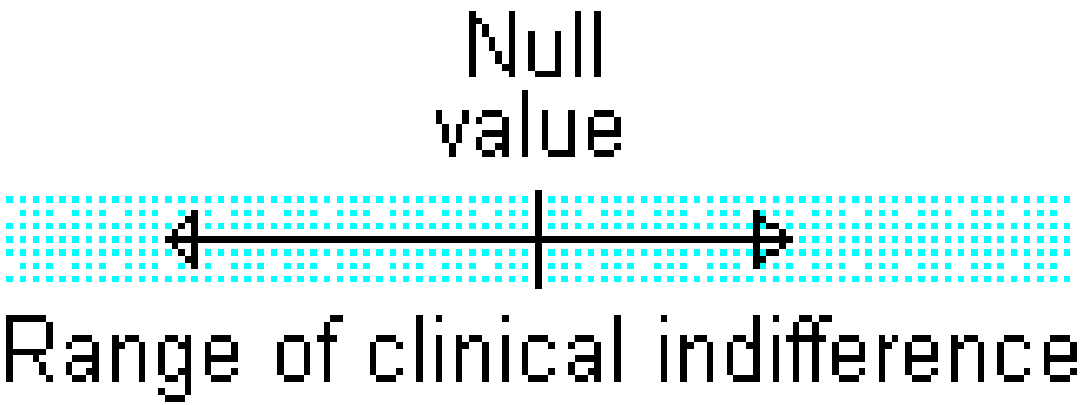


Figure 5: Interval entirely inside the range of clinical indifference

Speaker notes

You should also see whether the confidence interval lies partly or entirely within a range of clinical indifference. Clinical indifference represents values of such a trivial size that you would not want to change your current practice. For example, you would not recommend a special diet that showed a one year weight loss of only five pounds. You would not order a diagnostic test that had a predictive value of less than 50%.

Clinical indifference is a medical judgment, and not a statistical judgment. It depends on your knowledge of the range of possible treatments, their costs, and their side effects. As statistician, I can only speculate on what a range of clinical indifference is. I do want to emphasize, however, that if a confidence interval is contained entirely within your range of clinical indifference, then you have clear and convincing evidence to keep doing things the same way.

Confidence interval interpretation (5 of 7)

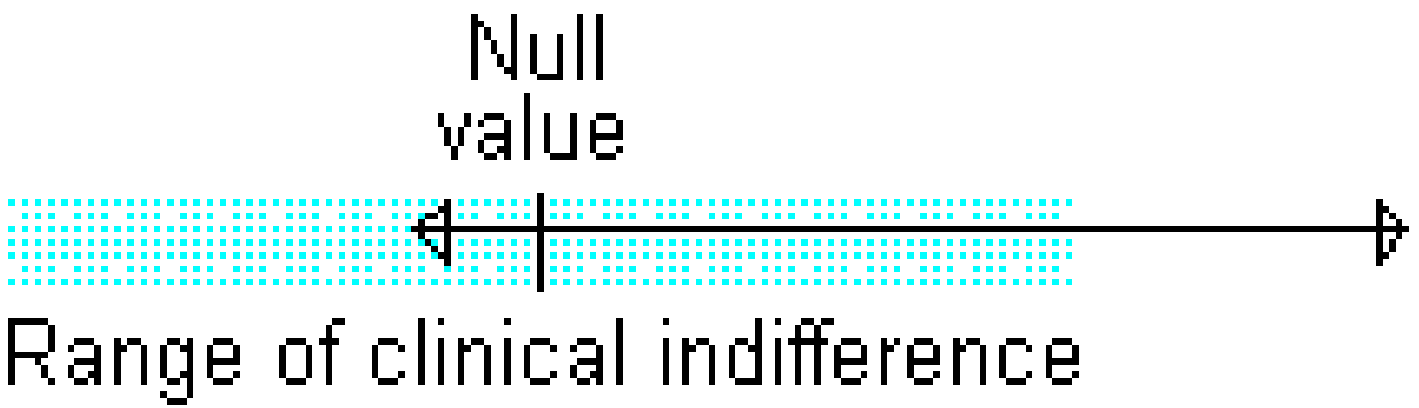


Figure 6: Interval partly inside/outside range of clinical indifference

Speaker notes

One the other hand, if part of the confidence interval lies outside the range of clinical indifference, then you should consider the possibility that the sample size is too small.

The interval contains zero, so it is plausible to behave as if the difference in population means or proportions is zero. But the interval also contains values that are clinically important. So it is plausible to behave as if there is a clinically important difference in means. How can you have two such different interpretations being plausible at the same time? That’s the definition of ambiguity. If you don’t like it, get used to it. Statistics will often identify areas of ambiguity, which is a good thing, because it tells us to not act prematurely, but instead demand more data before you make a definitive decision.

Quiz question, revisited

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. This confidence interval tells that the result is

1. statistically significant and clinically important.
2. not statistically significant, but is clinically important.
3. statistically significant, but not clinically important.
4. not statistically significant, and not clinically important.
5. The result is ambiguous.
6. I do not know the answer.

Speaker notes

Let’s go back to that question I posed earlier.

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

Well, it tells you that a relative risk of 1 (equal risks) is plausible, but that a relative risk of 2 (a doubling of risk) is also plausible. A tripling of risk is plausible. Good grief! This is an ambiguous result.

Doesn’t this bother you? It should. Someone ran a terrible experiment. An experiment so poorly designed that it can’t distinguish between no change in risk, or a tripling of risk.

It’s a terrible thing, but it happens all the time and it doesn’t seem to bother anyone but me. This is wretched. You got a hundred patients to let you poke and prod them. They took some bitter pills or maybe placebos. They are sacrificing their bodies in the name of science. And the best you can do is a confidence interval that goes from 0.82 to 3.94. Hang your head in shame!

Confidence interval interpretation (6 of 7)

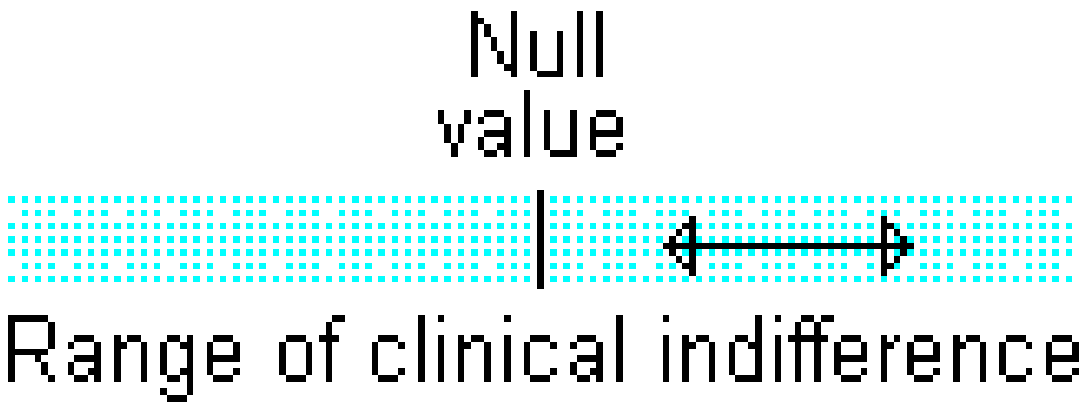


Figure 7: Confidence interval that contains the null value

Speaker notes

Some studies have sample sizes that are so large that even trivial differences are declared statistically significant, especially in this era of big data. If your confidence interval excludes the null value but still lies entirely within the range of clinical indifference, then you have a result with statistical significance, but no practical significance.

Confidence interval interpretation (7 of 7)

Figure 8: Confidence interval entirely outside the range of clinical indifference

Speaker notes

Speaker notes

Finally, if your confidence interval excludes the null value and lies outside the range of clinical indifference, then you have both statistical and practical significance.

Let’s talk about one more case. I don’t have a picture, but imagine a confidence interval that is mostly in the white region, the region of clinical importance, but the lower limit stretches into the range of clinical indifference. It doesn’t quite include the null value, but it comes within kissing distance. That’s a result that achieves statistical significance, but it does not provide definitive proof of clinical importance. No one ever talks about this case, but they should. Your confidence interval indicates statistical significance, but just barely. So don’t pretend that your results are the final word. You should not stop researching until you get a confidence interval that lies entirely inside or entirely outside the range of clinical indifference.

Why you might prefer a confidence interval

- Provides same information as p-value,
  - Clinical importance
  - Distinguish between
    - definitive negative result, or
    - more research is needed

What sample size do you need?

- Sackett’s formula
- Rules of thumb
- Confidence interval width
- Power calculations
- Post hoc power - never!
- Effect sizes - never!

Speaker notes

Speaker notes

I want to start the discussion of sample size with a formula proposed by David Sackett. It isn’t a deep mathematical formula, but rather something that tries to illustrate the intuition behind sample size calculations.

Then I wanted to present two useful rules of thumb, formal calculations involving confidence intervals and power. I also want to warn you against two very bad approaches, post hoc power and effect sizes.

cmaj jamc
CANADIAN MEDICAL ASSOCIATION JOURNAL JOURNAL DE L'ASSOCIATION MÉDICALE CANADIENNE

CMAJ. 2001 Oct 30; 165(9): 1226–1237.

PMCID: PMC81587
PMID: 11706914

Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!)

David L. Sackett

Figure 9: Sackett 2001, PMID:

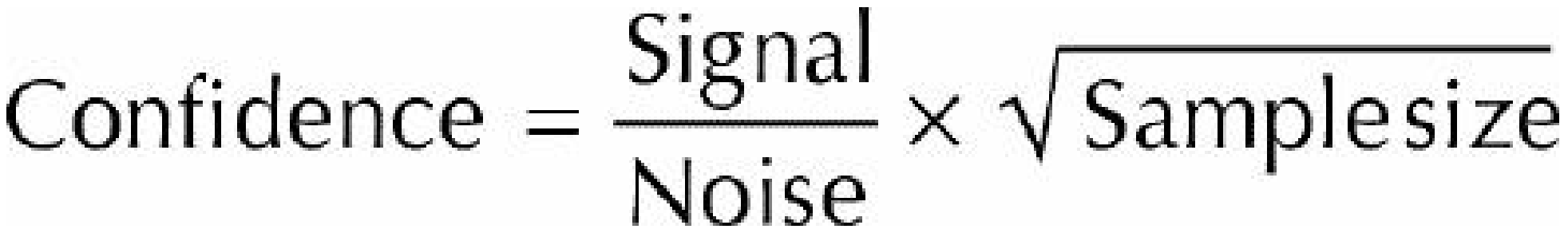


Figure 10: Formula found in Sackett 2001

Rules of thumb

- Rule of 50
  - Only for binary outcomes
  - Total sample size is irrelevant
  - Strive for 25/50 **events** in each group
- Rule of 16
  - $ES = MCD/\sigma$
  - $n = 16/ES^2$

Speaker notes

The first of these is the rule of 50. If you are measuring a binary outcome, you need to plan for a large enough sample size so that you see 25 to 50 events in each group. This is a very useful rule for seeing if your sample size needs to be in the hundreds, or if it needs to be in the thousands.

I was working with a researcher who wanted to study infants who were breastfeeding while at the hospital, go home and then come back in seven days with breastfeeding jaundice. I said, okay, about how often does this happen. And he said maybe somewhere between a half a percent and two percent of the babies. So I did the math. If the rate is on the high end, 2%, that means one out of every 50 patients will be readmitted. So you needs to get 50 times as many babies to assure that you will get 50 events. That's 2,500. In one group. And with the numbers on the high end of his interval. So I tell him, you'll probably need to study around 5,000 infants. He was not happy.

No one every got thrown in jail for violating a rule of thumb. But when you do the formal sample size formula, you sure as heck won't get a sample size of 100. A sample of 100 patients would mean that you'd expect to see two infants readmitted. Do you want to hang your hat on two infants?

The other rule, called the rule of 16, is useful for continuous outcomes. If you believe that a clinically important difference is a certain size, and that size is x% of a standard deviation, then you need to have 16/x-squared patients in each group. So if you believe that the clinically important difference is half a standard deviation, divide 16 by 0.5 squared to get 64 patients per group. If the clinically important difference is only one tenth of a standard deviation, watch out. You need to collect 16 divided by 0.1 squared or 1600 patients per group.

Confidence interval width

- How narrow do you want you confidence interval?
  - Algebraic solution
    - $\pm t_{0.975} SE = \pm MCD$
    - $SE = S_p \sqrt{1/n1 + 1/n2}$
    - Solve for n1 and n2
    - Usually assume n1 = n2
  - Trial and error

Speaker notes

Speaker notes

You can also justify your sample size by specifying how narrow you'd like your confidence intervals to be. This is especially useful in settings where there is no formal research hypothesis. The calculations here are a bit tedious, but manageable.

If you are not good with algebra, just try computing confidence intervals for various sample sizes until you stumble onto one that you like: not too wide and not too narrow.

Power calculations

- Need to specify MCD
- Power = $P[RejectH_0 \mid MCD] = 0.9$
- = $P[\bar{X}_1 - \bar{X}_2 > t_{0.975} SE \mid MCD] = 0.9$
  - Solve for n1 and n2.

Formal software for sample size calculations

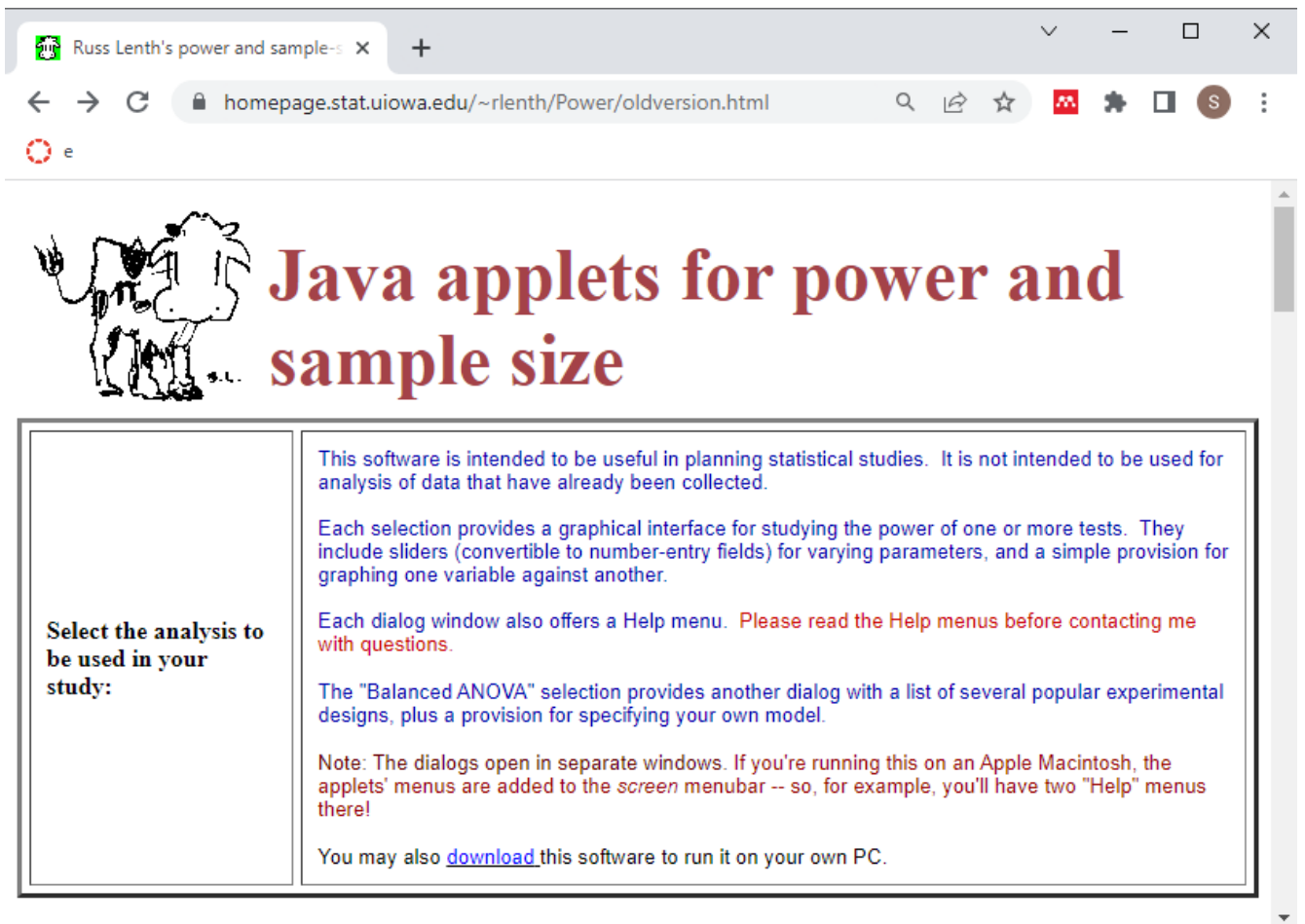


Figure 11: Lenth Power and Sample Size software

Post hoc power, effect sizes - never!

- Power must be calculated prior to data collection
- Effect sizes are only an intermediate calculation
  - Do not reflect clinical judgement
  - Always start with MCD
- Effect sizes also are not a useful summary statistic

Speaker notes

Speaker notes

Never try to justify your sample size after the fact. It is called a post hoc power calculation. It is very bad. If you didn’t have the foresight to justify your sample size before collecting your data, rely on the width of your confidence intervals to show that your sample size was adequate.

Your book and a lot of other books talk about effect sizes. While an effect size is a useful intermediate calculation, you cannot use the effect size, by itself, as it is impossible to interpret. Imagine a store with a sign in front announcing “Giant sale: all prices reduced by half a standard deviation”.

If you want to define a minimum clinically important difference, you must do it in the original scale of measurement, not using some unitless quantity like an effect size.

Criticisms of hypothesis testing (1 of 4)

- Criticisms of the binary hypothesis
  - Dichotomy is simplistic
  - Point null is never true
  - Cannot prove the null
- Possible remedy
  - $H_0 \ -\Delta \leq \ \mu_1 - \mu_2 \leq \Delta$

Speaker notes

There are many criticisms of hypothesis testing. You need to be aware of these criticisms, but I am not suggesting that you abandon hypothesis testing because of these criticisms.

The first set of criticisms deals with the binary nature of the hypotheses.

I've said many times that all dichotomies are false dichotomies, and I still hold to that. Hypothesis testing is a double dichotomy. You specify only two hypotheses, and you only two choices are accepting the null hypothesis and rejecting the null hypothesis. Shouldn't there be more than two choices?

Let me give an example. I found a couple of articles that studied the safety of vaccines. Now, vaccines are complicated, but let's try to understand the safety issue more clearly. It depends on the benefits of the vaccination combined with the probability that an individual will receive the benefit. Balance that against the harm caused by the vaccine combined with the probability that an individual will experience the harm. So how much harm and how probable does it need to be before you can say that you have a clinically important difference? Complicated, yes, but let's throw in a curve ball. How much do the harms and the probabilities need to be before you warn someone about those harms. How much do the harms and the probabilities need to be before you decide that you shouldn't be using this vaccine? Two very different questions, and two very different thresholds. So why do we force both of them into a single decision point?

Why shouldn't you allow a gray decision? So you could accept the null hypothesis for some values of the sample statistics, reject it for other values, and choose to neither accept nor reject for values intermediate.

Another thing about hypotheses is that a difference of exactly zero never actually occurs in the population. There's no way that, if you averaged a population of all males with particular disease and you averaged a population of all the females with a particular disease that you'd get the two to be exactly the same, even for a disease that has no association with gender. So what is the point of setting up a hypothesis and making a decision about it when you know in your heart of hearts that the null hypothesis is never true.

The other issue with hypothesis testing is that it does not allow you to prove the null hypothesis. If you really wanted to prove the null hypothesis, you have to do all sort of messy gyrations. Wouldn't it be nice to be able to act with the same level of certainty when you accept the null hypothesis as you do when you reject the null hypothesis? The very phrasing that some people use (fail to reject the null hypothesis in place of accept the null hypothesis) shows how convoluted things get.

Many of these criticisms of the binary hypothesis would disappear if we allowed for testing not equality in the null hypothesis but rather testing whether the difference is within some interval. But this is not going to happen anytime soon.

Criticisms of hypothesis testing (2 of 4)

- Criticisms of the p-value
  - Not intuitive, easily misunderstood
  - “results more extreme”
  - Ignores clinical importance
  - Does not measure uncontrolled biases

Speaker notes

Speaker notes

The p-value that lies at the heart of hypothesis testing has also been roundly criticized. It seems backwards in more than one way. It is evidence against a hypothesis rather than for a hypothesis. You get lots of evidence against the null hypothesis when the p-value is small and little or no evidence when the p-value is large. And the conditional probability that your p-value represents, the probability of getting sample results or results more extreme given that the null hypothesis is true represents the reverse of what you really want. What you really want is the probability that the null hypothesis is true given the data that you observed.

A lot of people dislike the idea of looking for a probability involving the sample results or results more extreme. Why, they ask, do you want a probability involving results more extreme. You didn’t observe a value more extreme. You observed a single value.

The p-value ignores clinical importance. This would be easy to fix if people used an interval -delta to delta for the null hypothesis as mentioned earlier, but no one is ready to do this.

Finally, the p-value is unaffected by threats to internal validity. If you conduct the study poorly, such as failing to keep information away from patients in a blinded study, or having a large number of protocol violations, that should be reflected in your p-value. But the p-value ignore these problems.

Criticisms of hypothesis testing (3 of 4)

- General criticisms
  - Too hard to reject H0
  - Too easy to reject H0
  - Too reliant on a single study
  - Thoughtless application

Speaker notes

Hypothesis testing has been criticized because it is too hard to reject the null hypothesis, especially for small samples in a noisy setting. The conservative inside of you and me probably thinks this is good. Don't make a choice between two therapies or drugs until you've accumulated sufficient evidence. But failing to act quickly can sometimes force you to pay a price. This relates to the trade-offs that you see all the time between false negative results and false positive results. The hypothesis testing framework is biased towards preventing a false positive result (a Type I error) and it is very difficult to get this framework to work well when a false negative result is worse. This can occur, for example, in a setting with a 100% fatal disease with no known cure.

The opposite problem is also true. It is often too easy to reject the null hypothesis. In this era of big data, you can quickly get millions of data points or more and you have so much precision that any sample statistics even of a trivial size will lead to a statistically significant result. This is not good. Instead of identifying one or two risk factors as being statistically significant, a really big data set will identify hundreds or even thousands of risk factors as being statistically significant.

The p-value is too reliant on a single study and does not consider what previous research has been done.

My biggest criticism of p-values, though, is the thoughtless way in which they are applied. I've written about the p-value receptor inside a scientist's brain.

A research team took ten scientists and placed them inside an fMRI. The fMRI shows which parts of your brain are active as your brain processes different types of information. The scientists were shown a variety of graphs taken from actual peer-reviewed publications.

As you might expect, the part of your brain that activates first when you are presented an image of a graph is your visual cortex. For most graphs, this was quickly followed by an activation of the parietal lobe, the part of your brain responsible for numerical computations.

But some graphs showed a different pattern. If the graph included a p-value, activation of the visual cortex is followed by activation of an area of your amygdala that is as yet poorly understood. The research team called this portion of the amygdala the p-value receptor.

If your p-value receptor is activated and the p-value is larger than 0.05, the p-value receptor sends strong signals to the pain centers of the brain. This is clearly an adaptive behavior. Scientists who routinely produce p-values larger than 0.05 will not survive and reproduce.

If the p-value receptor is activated and the p-value is 0.05 or smaller, the p-value receptor sends strong signals to the pleasure centers of the brain. Again this is an adaptive behavior. But the interesting finding is that there is a dose response effect. The p-value receptor produces about the same level of pleasure stimulation for p-values of 0.05, 0.04, 0.03, and 0.02. But p-values of 0.01 show an increase in

stimulation that becomes even strong for p-values of 0.0099 and smaller. Perhaps there is some pattern associated with p-values that have two zeros to the right of the decimal place that is stronger then a p-value with just a single zero.

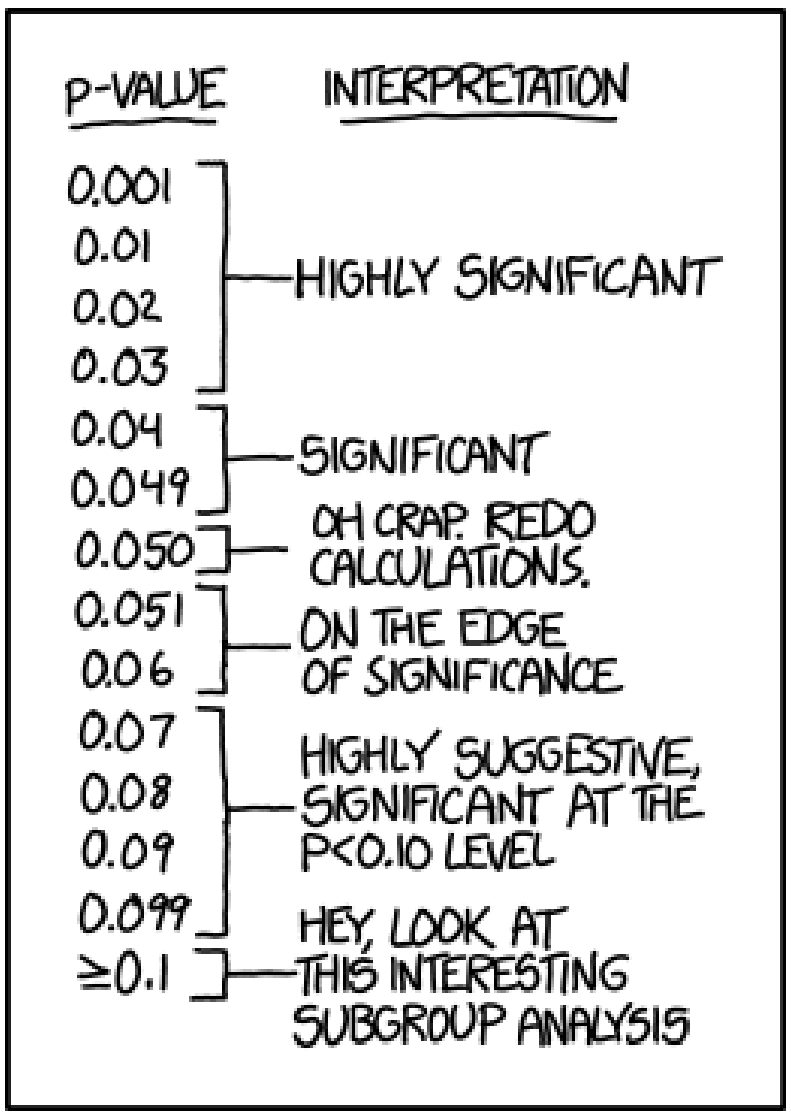
The scientists also examined the effect of p-values reported in scientific notation. There was an increase in latency when the p-value receptor is fed a p-value in scientific notation. This probably represents an attempt to decode the scientific notation. But p-values in scientific notation with exponents of -4 or smaller showed an eventual spike in activation of the pleasure centers of the brain that are comparable to those achieved during orgasm.

The research also noted a second important effect of the p-value receptor. Once the p-value receptor is stimulated, the entire cerebral cortex, the portion of your brain associated with logic and complex thinking, is immediately shut down. This insures that a scientist's brain will focus only on the pleasure or pain associated with the p-value and will ignore the power of the study, the magnitude of the treatment effect, and other unimportant issues.

The researchers suggest that statisticians who want to earn more consulting income and insure repeated business should do their best to produce only p-values that stimulate the pleasure centers of the brain.

This is from my blog: http://blog.pmean.com/scientist-brain/

Criticisms of hypothesis testing (4 of 4)



Speaker notes

I have to end with another cartoon. Both of these cartoons were drawn by Scott Munro, the creator of the xkcd comic series.

This is a cartoon showing a table of p-values with various labels.

0.001, 0.01, 0.02, and 0.03 are labeled “Highly Significant”.

0.04 and 0.049 are labeled “Significant”.

0.050 is labeled “Oh crap. Redo calculations.” This is because no one knows what exactly to do with a p-value that is right on the boundary.

0.051, 0.06 are labeled “On the edge of significance”. If these values are on the edge of significance then the 0.04 and 0.049 should share that same label. But no one uses these hedging terms unless they are on the “wrong” side of 0.05.

0.07, 0.08, 0.09, 0.099 are labeled “Highly suggestive, relevant at the $p<0.10$ level”. I personally don’t have a complaint here, but others consider this a post hoc modification and a form of p-hacking.

≥0.1 is labeled “Hey, look at this interesting subgroup analysis”. This is a reference to p-hacking. If your primary p-value is not statistically significant, hunt for some other p-values.

What should you do if you do not have a hypothesis to test?

- Descriptive statistics
  - Include confidence intervals
- Qualitative data analysis

Bayesian example

Teaching Inference about Proportions Using Bayes and Discrete Models

Jim Albert
Bowling Green State University

Journal of Statistics Education v.3, n.3 (1995)

Albert 1995

Speaker notes

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. I’m going to modify things just a little bit.

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

- $P(H|E) = P(E|H)P(H)/P(E)$
  - H = hypothesis
  - E = evidence (data)

Prior

- P[H] is prior
  - Subjective prior
    - Contrast optimistic/pessimistic perspectives
    - Incorporate prior knowledge
  - Flat (non-informative 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.

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.

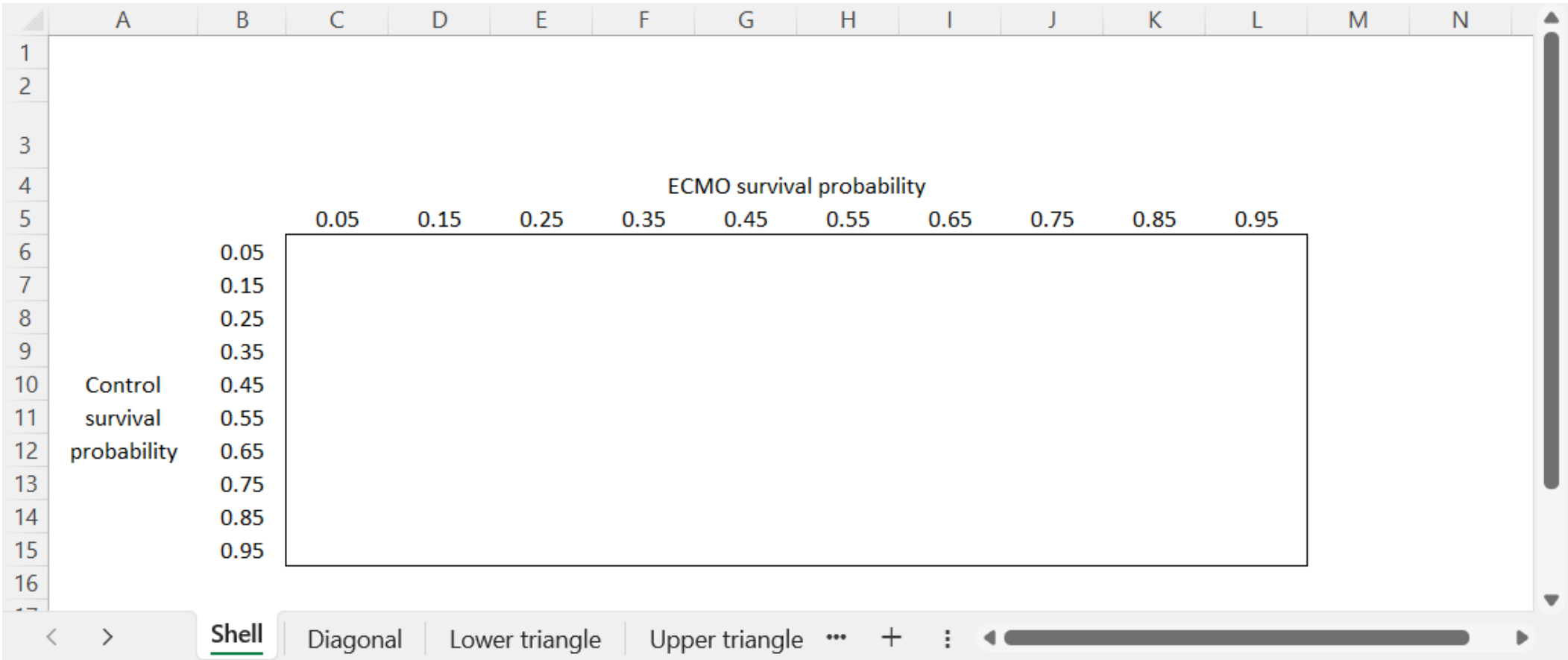


Figure 13: Empty table for prior probabilities

Speaker notes

Speaker notes

First layout the hyptohesized survival probabilities. The columns represent hypothesized probabilities of survival for ECMO and the rows represent hypothesized probabilities for conventional therapy.

Now if I was doing this seriously, I'd list a few hundred or thousand probabilities across the top and down the side. But that wouldn't fit on a single slide.

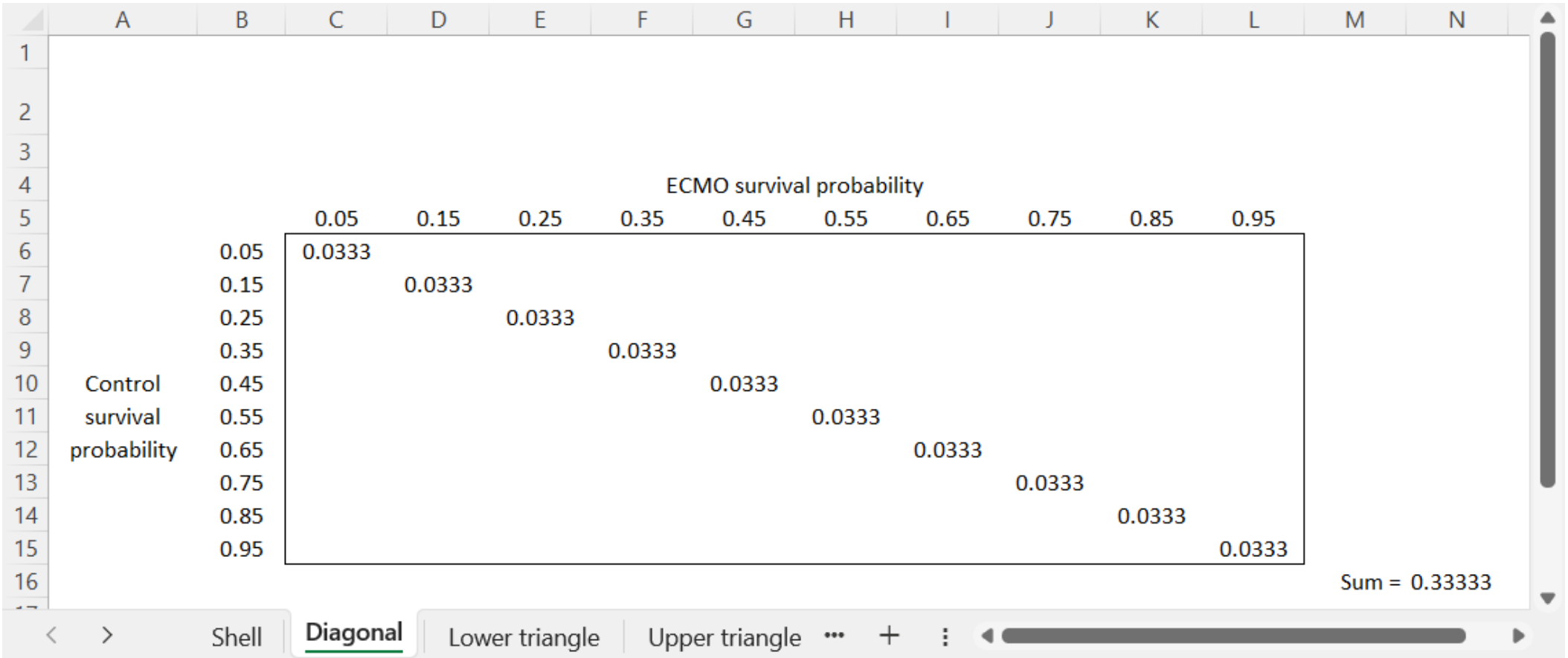
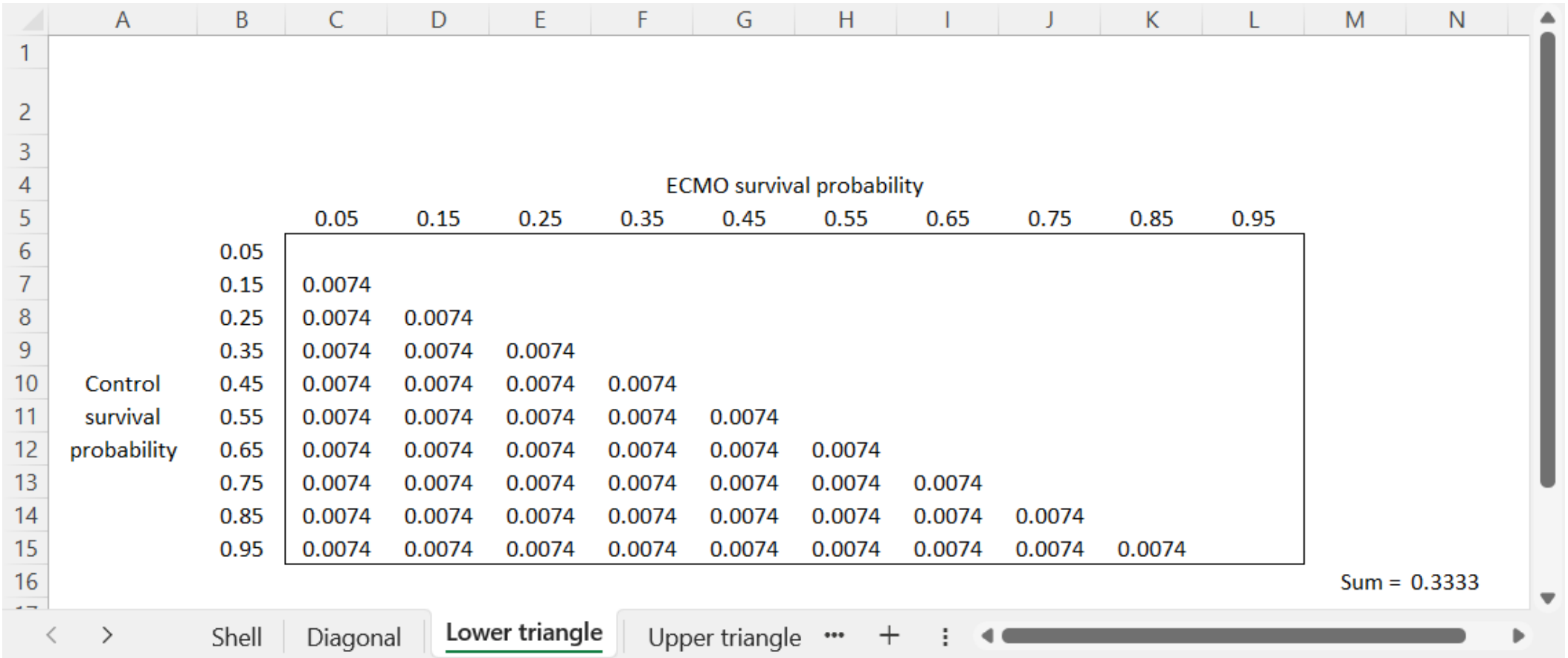


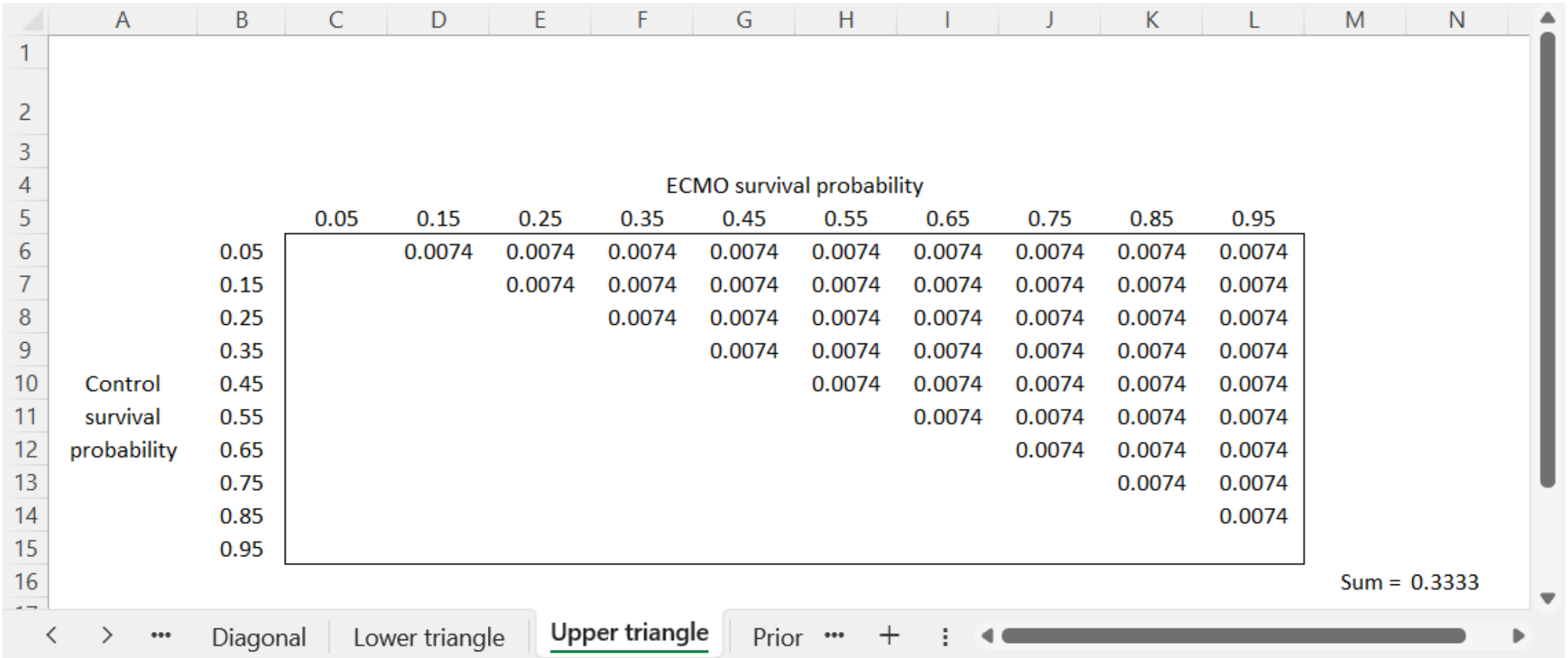
Figure 14: Table with diagonal priors



		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										
	0.15	0.0074									
	0.25	0.0074	0.0074								
	0.35	0.0074	0.0074	0.0074							
	0.45	0.0074	0.0074	0.0074	0.0074						
	0.55	0.0074	0.0074	0.0074	0.0074	0.0074					
	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

Figure 15: 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
	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
Control	0.45						0.0074	0.0074	0.0074	0.0074	0.0074
survival	0.55							0.0074	0.0074	0.0074	0.0074
probability	0.65								0.0074	0.0074	0.0074
	0.75									0.0074	0.0074
	0.85										0.0074
	0.95										

Sum = 0.3333

Figure 16: Table with upper triangle of prior probabilities

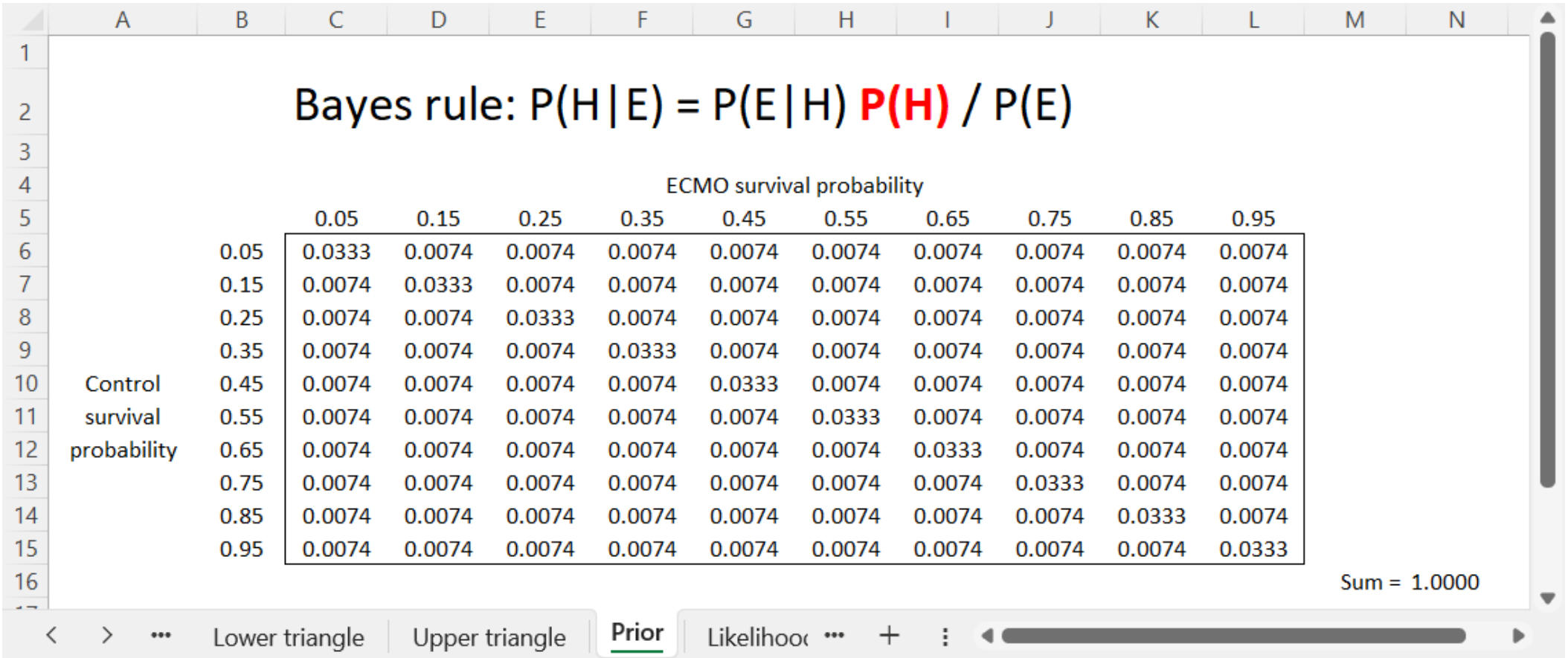
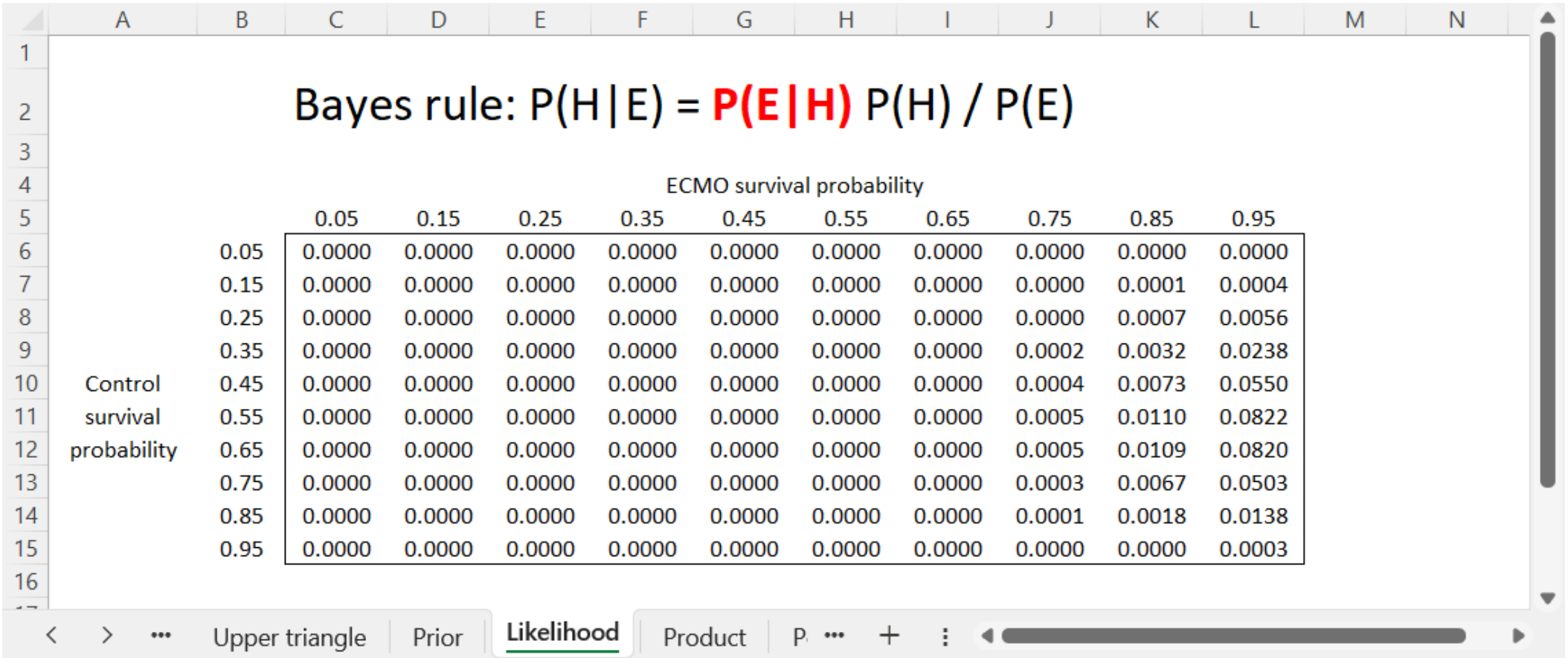


Figure 17: Complete table of prior probabilities



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

Figure 18: Table of likelihoods

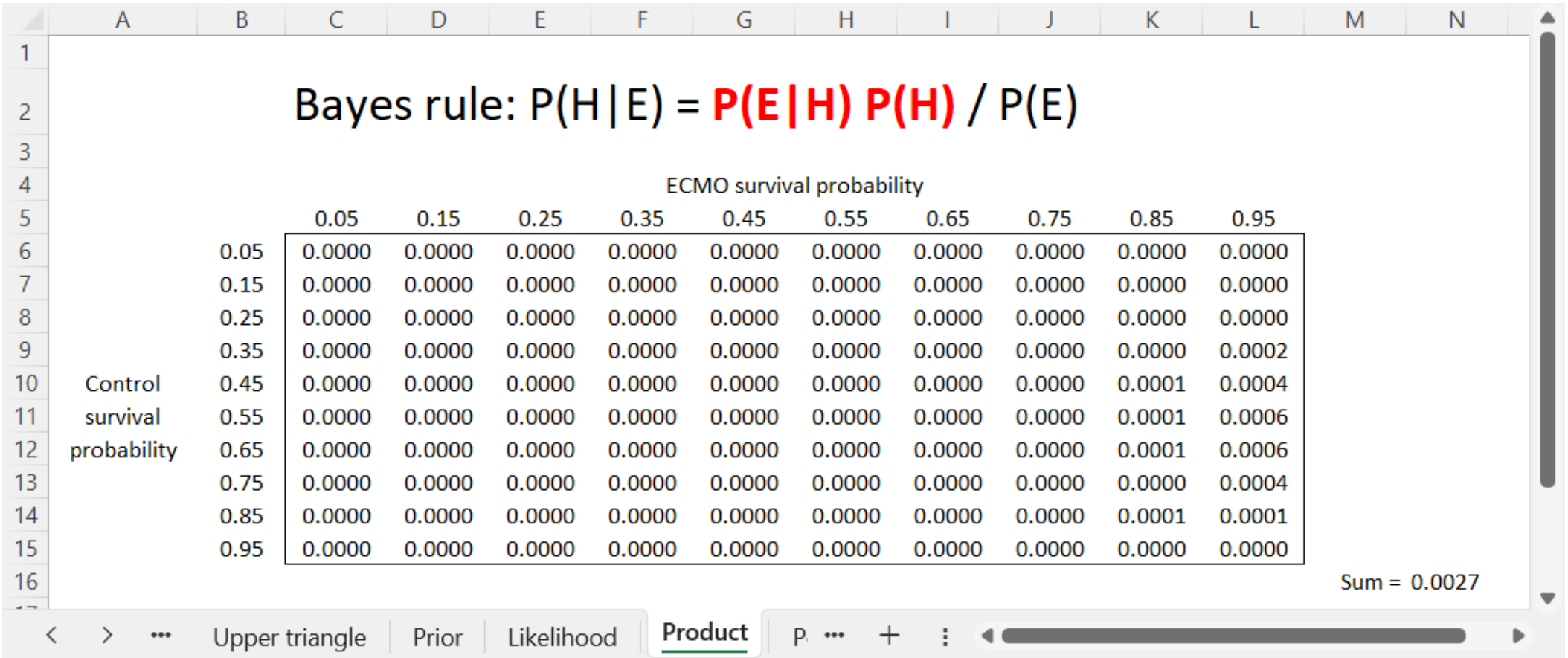
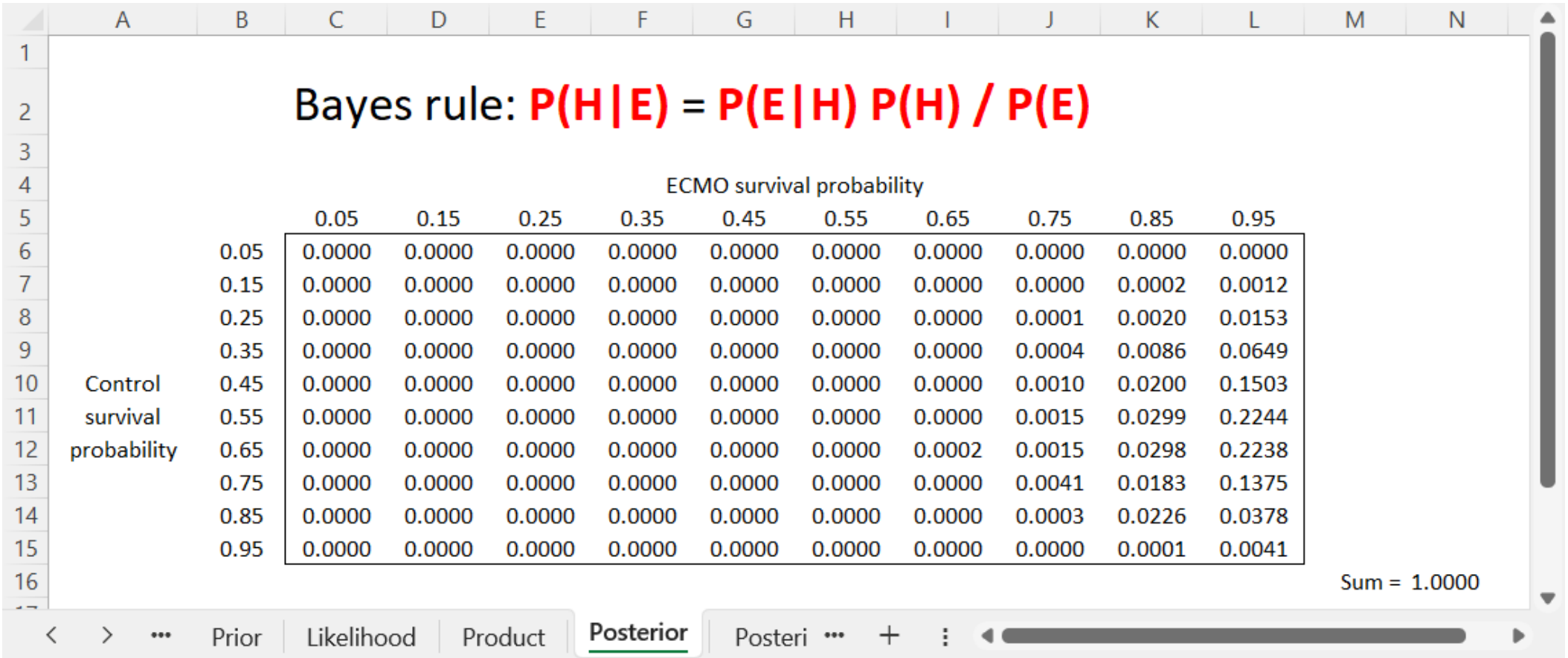


Figure 19: Product of prior and likelihood

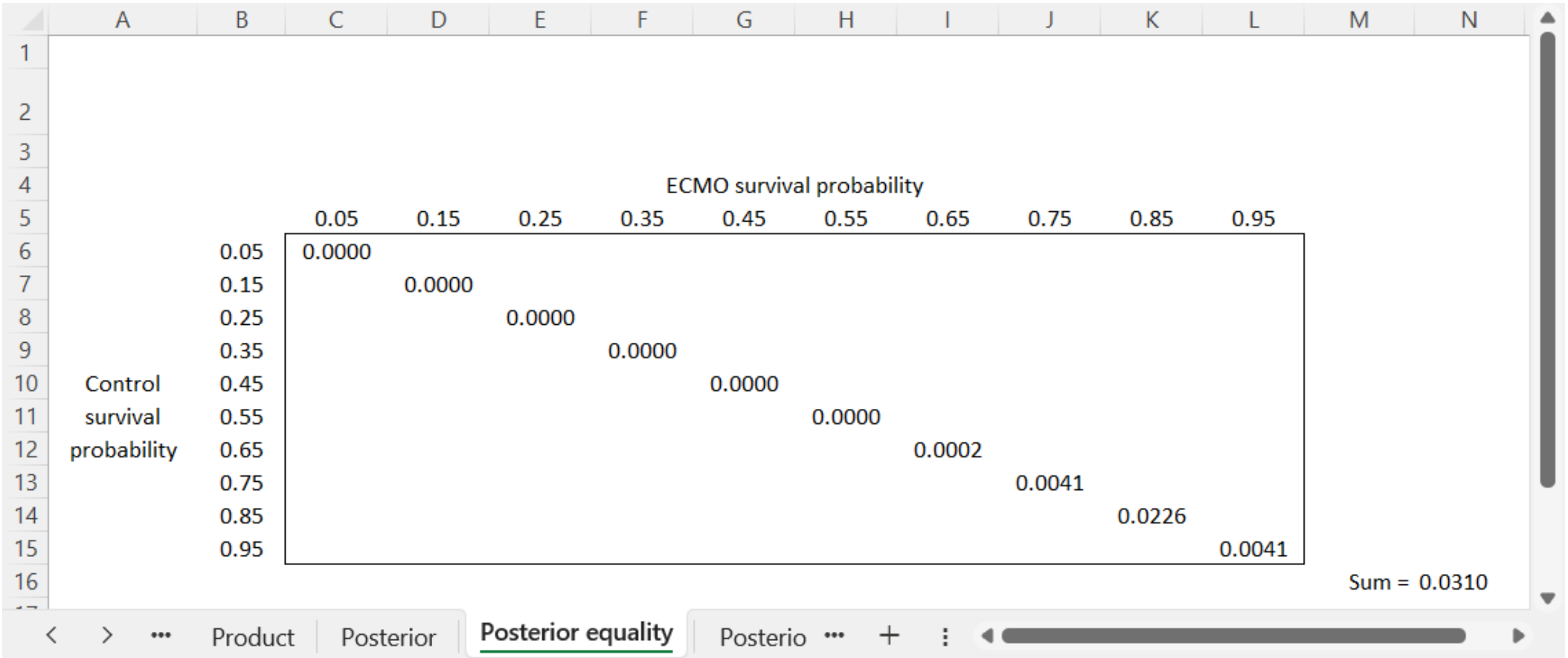


Bayes rule: $P(H|E) = P(E|H)\,P(H) / P(E)$

		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	0.0002	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

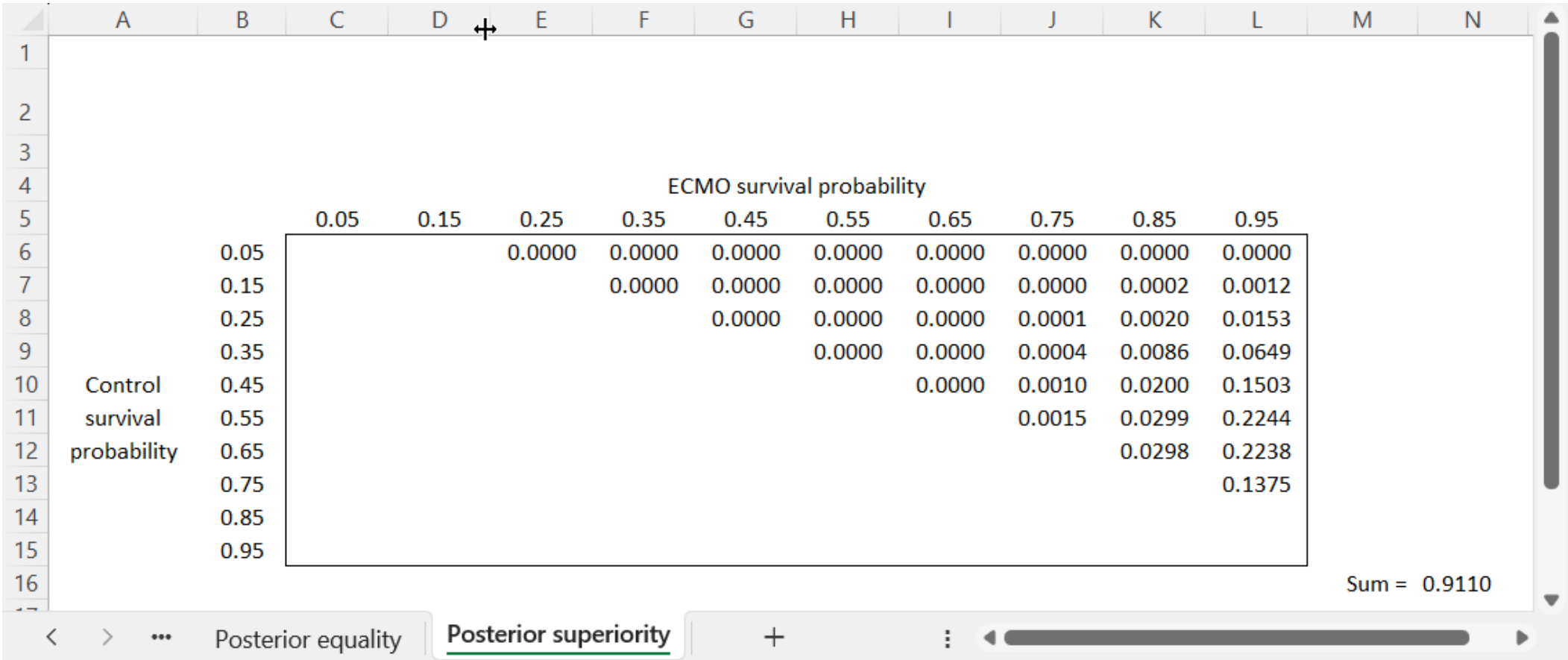
Figure 20: Table of posterior probabilities



		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	0.0000									
	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 21: 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
	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
Control	0.45							0.0000	0.0010	0.0200	0.1503
survival	0.55								0.0015	0.0299	0.2244
probability	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

Speaker notes

Speaker notes

Posterior probability of 20% or greater superiority

Criticisms of Bayesian data analysis

- Choice of prior distribution is arbitrary
- Probability of a hypothesis is absurd
- Requires strong distributional assumptions
- Computationally intensive

Areas where Bayesian data analysis excel

- Imputation
- Latent models
- Random effects/hierarchical models
- Incorporating historical data

Repeat the bad quiz question

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

1. Strong evidence for the null hypothesis
2. Strong evidence for the alternative hypothesis
3. Little or no evidence for the null hypothesis
4. Little or no evidence for the alternative hypothesis
5. More than one answer above is correct.
6. I do not know the answer.

Speaker notes

Speaker notes

Here’s that bad quiz question again. Why are none of these answers correct?

WAIT FOR STUDENTS TO RESPOND

Because a p-value is a measure of evidence against the null hypothesis. A large p-value means little or no evidence against the null hypothesis.

Repat the bad confidence interval question.

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. This confidence interval tells you that the result is

Speaker notes

Speaker notes

What does this confidence interval say about the research?

WAIT FOR STUDENTS TO RESPOND

The result is ambiguous. You should be recoiling at this confidence interval because it cannot distinguish between no change in risk, a doubling of risk, or even a tripling of risk.

Repeat of Bayesian question

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

1. data shrinkage.
2. a prior distribution.
3. a posterior distribution.
4. p-values.
5. I do not know the answer.

Speaker notes

Here’s a good question. See if you remember a key feature of Bayesian data analysis.

Summary

In today's class, you learned about

- p-values,
- confidence intervals,
- justifying your sample size, and
- Bayesian data analysis

Are there any questions?

Speaker notes

You saw a lot today. I hope you are more comfortable with some of these concepts.