

## Using the Pachinkogram to introduce and explore Bayes' theorem

Dr Marie Fitch  
*The University of Auckland, New Zealand*  
Email: m.fitch@auckland.ac.nz

### Overview of lesson

A person has been randomly chosen to be tested for diabetes and tests positive. What is the probability they actually have diabetes? Students use the Pachinkogram visual tool to explore this scenario and investigate what happens when the given probabilities change. Rather than introducing Bayes' theorem using formulae, this lesson aims to give students an intuitive understanding of the nature and effect of given and resultant probabilities.

### Learning objectives

- Realise that, in general, inverting a conditional probability changes its value
- Be able to identify a specific situation when inverting the conditional probability does not change its value
- Be able to explain in broad terms how changing the base rate affects the value of a conditional probability
- See that sample probabilities vary around the true probability

### Suggested age range

First-year University

### Time required

One 50-minute lecture slot

### Keywords

conditional probability, Bayes' theorem, pachinkogram

## Introduction

After teaching our first-year university theoretical probability course ( $n \approx 150$  students) for a few years I realised that our typical mix of students ranged in both mathematical ability and confidence from very high to those that struggle. I watched the Pachinkogram tool being developed (Budgett & Pfannkuch, 2019) and I soon saw it had great potential for adding into the course.

Although the tool was primarily developed as an interactive visual tool to dispel common misconceptions, the base rate fallacy and confusion of the inverse, its appeal in this course went much deeper. Through using the tool, students see that  $\mathbb{P}(A|B)$  and  $\mathbb{P}(B|A)$  are different. They are then guided to begin thinking about what affects their relative values, culminating with considering whether or not they can ever be the same. Alongside that I have noticed that the tool, for those students who typically apply formulae with no understanding, results in them developing a deeper understanding of what they are doing when they use Bayes' theorem. For those students leery of formulae, the tool assists them in developing a more intuitive understanding of Bayes' theorem before the formula is introduced. Because the tool is simulation based, probabilities vary around given and calculated values, and hence a short part of the task uses this to remind students of sampling variability, something that is often lost in a theoretical probability course.

The lesson is structured as a guided exploration mainly due to time constraints. The worksheet that the students are given is attached at the end of this lesson plan.



Image: CreativeCommons/Wansolwara

## Lesson outline

### 1. Conjectures

I begin the lesson by asking the students to answer the following three questions at the top of the worksheet that they are given. I point out that I am not going to look at what they have written and that I want them to come back to their answers at the end of class and see if they would change any of them.

- Does  $\mathbb{P}(A|B) = \mathbb{P}(B|A)$  always / never / sometimes (when)?  
*Circle one*
- What affects the values of  $\mathbb{P}(A|B)$  and  $\mathbb{P}(B|A)$ ?
- Do sample estimates of conditional probabilities have the same value as the theoretical calculations?

The point of this step is to acknowledge prior learning and bring any misconceptions to the forefront. It also serves to give a learning focus to the students by following a typical learning strategy in probability, which is to make conjectures and then to test them out.

### 2. Scene setting

I chose diabetes as the context in the first place because this is the default labelling of the Pachinkogram branches! However, the added bonus in using this context is that it is a familiar context for most students. I begin the lesson by asking who has diabetes or knows someone (family member, friend, acquaintance) with diabetes. There is usually a large proportion of the class with raised hands. We then discuss the prevalence of diabetes in New Zealand, and the need for testing. We talk about a blood test being a relatively cheap and non-invasive test, but that the resulting classification (positive or negative for diabetes) is not always correct. The ideas behind and meanings of sensitivity and specificity (“let’s learn some medical jargon”) are introduced. Based on research on diabetes prevalence (Budgett & Pfannkuch, 2019), sensitivity and specificity, students are encouraged to fill-in the relevant probabilities on their worksheet.

It is estimated that in New Zealand roughly 7% of the population have type 2 diabetes. One of the many tests for diabetes has a sensitivity of about 85% and a specificity of about 88%.

This means that  $\mathbb{P}(\text{Positive test}|\text{Diabetic}) =$   
and  $\mathbb{P}(\text{Negative test}|\text{not Diabetic}) =$

I then ask them to imagine that it has been decided to screen all New Zealanders for diabetes, and a person tested positive. What would be the probability that they

had diabetes? Does the research tell us? That is, is it one of the probabilities that is given in the above information?

Initially I didn’t include a discussion at this step, but found that students didn’t know what they were doing or why. Because the students are used to lectures and having everything on the screen, I usually have slides prepared that explicitly give the prevalence, sensitivity and specificity by name and as probability statements that they can copy to their worksheets.

### 3. Exploration

The students are now encouraged to form into groups of 2 to 4 around a device and go to [stat.auckland.ac.nz/~vt/](http://stat.auckland.ac.nz/~vt/). I usually go to the website myself and demonstrate moving the sliders. Students are then largely left to independently work through the worksheet in their groups starting with questions 1 to 3.

Students have previously been asked to bring a laptop or tablet to class. I generally discourage the use of mobile phones as the small screen makes moving the sliders quite difficult. I also actively encourage working in groups so that students are verbalising their thinking and learning from each other.

Go to [www.stat.auckland.ac.nz/~vt/](http://www.stat.auckland.ac.nz/~vt/)

Adjust the sliders so that the percentages match those above.

1. When you sample where will most ‘people’ end up?  
*Try ‘testing 1000 people’ that is select sample once.*
2. What do you notice?
3. What is the estimate from this sample of  $\mathbb{P}(\text{Diabetic}|\text{Positive test})$ ?  
*Click Flip to see this*

Figures 1, 2 and 3 illustrate the three components of the Pachinkogram tool that the students interact with.

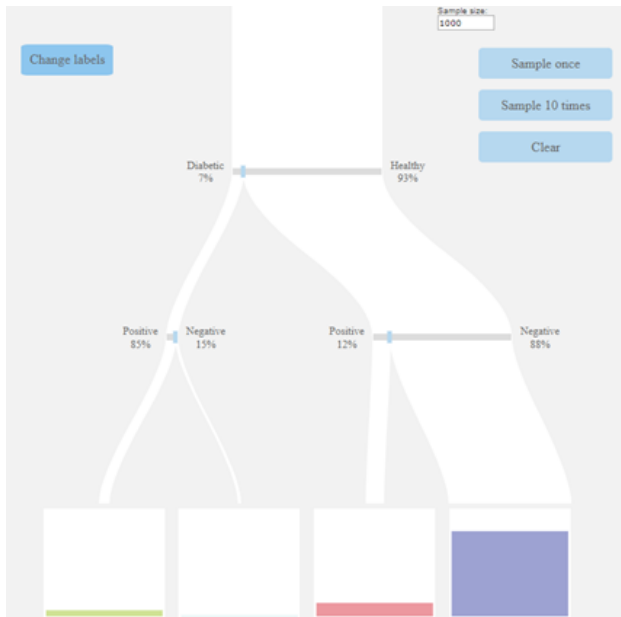


Figure 1: Pachinkogram for diabetes screening scenario after sample of size 1000 taken

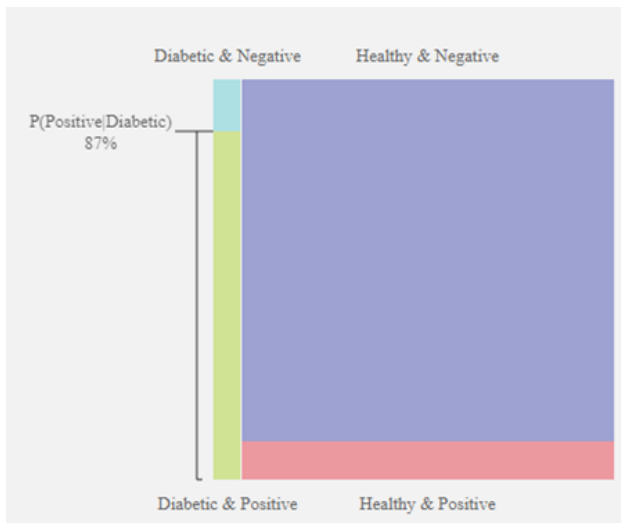


Figure 2: Eikosogram conditioned on diabetic status ( $n=1000$ )

Figure 1 shows a probability tree where the branches are proportional to the probability. When students take a sample, balls drop down the tree into one of the buckets at the end of each branch. The coloured rectangles in the buckets represent the proportion of balls in each bucket after sampling.

Figure 2 shows the Eikosogram which appears on the top right of their screen.

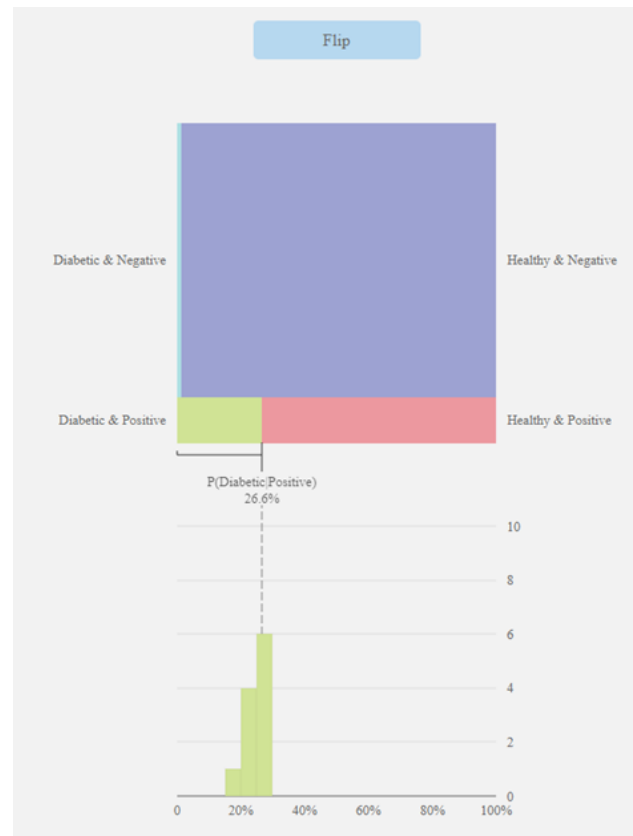


Figure 3: Eikosogram conditioned on screening test status with histogram showing  $P(\text{Diabetic}|\text{Positive})$  after sampling  $n=1000$  eleven times

My students have seen and worked with an Eikosogram in the previous lecture. If this is not the case you would need to explain an Eikosogram at this point (see Budgett & Fitch (2018) for further information).

Figure 2 is an example of what students see after taking one sample of size 1000. Each coloured rectangle has an area proportional to the area of the rectangle of the same colour at the bottom of the probability tree (Figure 1). The width of the green rectangle is 6% of the overall width of the square (i.e. the same as  $\mathbb{P}(\text{Diabetic})$ ). Students quickly see that  $\mathbb{P}(\text{Positive}|\text{Diabetic})$  is similar to but not the same as the probability they were given (sampling error).

When students click the flip button above the Eikosogram (Figure 3) the rectangles realign, keeping the same area, to show  $\mathbb{P}(\text{Diabetic}|\text{Positive})$ . When students click the 'Sample 10 times' button the Eikosogram defaults to the Figure 3 view. Figure 3 also shows the histogram below which illustrates the sampling variability.

I move around the class, interacting with the groups answering their questions and asking questions to focus and provoke thought. Often, I pull the class back together to discuss findings part way through, for example, to discuss questions 4 and 5. At this point students are generally pretty clear that  $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$ . Many are not clear exactly why but are beginning to get the idea that the prevalence seems to affect things.

4. Was this sample unusual?  
*Select sample 10 times to check. What do you notice?*  
 Why is  $\mathbb{P}(\text{Diabetic}|\text{Positive test})$  so small?  
 5. Some sub-populations have a higher prevalence of diabetes.  
*Clear your results and change the prevalence from 6% to 12%.*  
 Does  $\mathbb{P}(\text{Diabetic}|\text{Positive test})$  change? How?

I then encourage them to try the different scenarios in questions 6 and 7.

6. If you were being tested for diabetes what would you like  $\mathbb{P}(\text{Diabetic}|\text{Positive test})$  to be?  
*Try adjusting the prevalence.*  
 What prevalence would be needed for this test to be correct with the probability you desire?  
 7. Other tests have a higher sensitivity and specificity.  
*Change both the sensitivity and specificity to 99%.*  
 What effect does this have on the prevalence needed for you to get your desired  $\mathbb{P}(\text{Diabetic}|\text{Positive test})$ ?  
 Is it still possible for  $\mathbb{P}(\text{Diabetic}|\text{Positive test})$  to be small?

Student comments suggest that the explorations at this point help their understanding: “The model was effective in illustrating the role that prevalence has on conditional probabilities.” “It’s helpful since it shows a visible diagram which can be altered for different result, which made it easier to think about the outcome and how it occurred.”

Having thoroughly convinced the students that  $\mathbb{P}(A|B)$  and  $\mathbb{P}(B|A)$  are different, and that the prevalence affects how different they are, the final part of the exploration, which not all students get to in class, is a challenge.

8. Is it ever possible for  $\mathbb{P}(A|B) = \mathbb{P}(B|A)$ ? Try it!

The point here is to get students thinking about what is going on and trying different scenarios to try and achieve equality. The most common strategies are to increase the prevalence to almost or exactly 1 or to make the prevalence, sensitivity and specificity all 0.5. In their next lesson, they will meet Bayes’ theorem and see mathematically how this could occur.

#### 4. Wrap-up

As the end of the lecture time approaches I pull the class back together and ask the students to go back to the initial three questions and compare their conjectures with the answers derived from the tool. Is there anything they would change? If so, note it at question 9.

9. Refer back to the questions in the box at the top of the first page ... Summarise what you have learnt 😊

If time allows, we also have a general discussion of what they have discovered and learned. Although I do not mention the base rate fallacy or confusion of the inverse by name (but maybe should) I always make sure that in my discussions with individual groups, and in this final summary, that both misconceptions are discussed and ‘corrected’. I also point out that at the next lecture we will be considering a mathematical formula, Bayes’ theorem, for finding  $\mathbb{P}(B|A)$  when we have been given  $\mathbb{P}(A|B)$ . The lesson finishes with an encouragement to go home and “try more scenarios”.

Student comments suggest that overall this activity allows them to focus on what is occurring, rather than getting bogged down in calculations and also that it increases their understanding of conditional probabilities. “A good way to understand how conditional probabilities worked without having to perform calculations at each step” “It is a useful tool for visualising conditional probabilities and it has improved my understanding.” “It helped me understand better about how conditional probability works by visualising the process.” “It gets the calculating process visualized while solving questions about conditional possibilities and it was way more obvious to find out all possibilities involved in questions. Moreover, it seems to be able to strengthen my probabilistic reasoning ability.”

#### Adaptations

It is relatively simple to extend this activity by getting students to connect what they are finding in the tool with mathematics. The more able students can be challenged to essentially ‘discover’ Bayes’ theorem for themselves. Mathematically weaker students can be encouraged to focus just on the tool and gain confidence in seeing what is occurring and hence begin to develop an intuition for what to expect.

Student comments suggested possible extensions in terms of linking the tools with mathematical formulae. “The Pachinkogram was quite helpful in terms of visualising the data in a probability tree and seeing what part of the tree the conditional probability represented. However, it does not represent the equation for conditional probability very well.” (It) “would be useful to have the option to add more than two levels of condition as well as more than two partitions per level.”

### Teacher notes

The references I have included below could be split into two groups. One group consists of three articles that focus on the confusion of the inverse fallacy and using the Pachinkogram. In particular the conference paper co-written by myself and Stephanie Budgett, describes in more detail the background to this lesson, including a brief explanation of an Eikosogram. The other group comprises four articles that talk about the worth of using visualisations and active learning. They provide some helpful background reading.

Finally, I include here the simplest version of Bayes’ theorem for quick reference.

$$\text{Bayes' theorem: } \mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)}$$

### References

Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics*, 52, 215–241.

Budgett, S. & Fitch, M., (2018). Conditional probability, friend or foe? In M.A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan*. Voorburg, The Netherlands: International Statistical Institute. [https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10\\_6A1.pdf](https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_6A1.pdf)

Budgett, S., & Pfannkuch, M. (2019). Visualizing chance: Tackling conditional probability misconceptions. In G. Burrill & D. Ben-Zvi (Eds.), *Topics and trends in current statistics education research: International perspectives* (pp. 3–25). Cham, Switzerland: Springer Nature Switzerland AG. [https://doi.org/10.1007/978-3-030-03472-6\\_1](https://doi.org/10.1007/978-3-030-03472-6_1)

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415.

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and patients. *Social Science and Medicine*, 83, 27–33.

Gigerenzer, G. (2014). *Risk savvy: How to make good decisions*. New York, NY: Viking.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes’ theorem and the additivity principle. *Memory & Cognition*, 30, 171–178. doi:10.3758/BF03195278

### Materials required

- One web-enabled device (preferably laptop or tablet) per three students
- One handout for each student

### Copyright information

Authors maintain copyright of their published material in *Statistics and Data Science Educator*. Any person requesting permission to use materials from a *Statistics and Data Science Educator* lesson in a publication must obtain permission from the authors of the lesson.

## Pachinkogram and conditional probabilities

- \* Does  $\mathbb{P}(A|B) = \mathbb{P}(B|A)$  always? never? sometimes? (when?)
- \* What affects the values of  $\mathbb{P}(A|B)$  and  $\mathbb{P}(B|A)$ ?
- \* Do sample estimates of conditional probabilities have the same value as the theoretical calculations?

It is estimated that in New Zealand roughly 6% of the populations have type 2 diabetes. One of the many tests for diabetes has a sensitivity of about 85% and a specificity of about 88%.

This means that  $\mathbb{P}(\text{Positive test}|\text{Diabetic}) =$

and  $\mathbb{P}(\text{Negative test}|\text{not Diabetic}) =$

Go to [www.stat.auckland.ac.nz/~vt](http://www.stat.auckland.ac.nz/~vt)

Adjust the sliders so that the percentages match those above.

1. When you sample where will most 'people' end up?

*Try 'testing 1000 people' that is select sample once.*

2. What do you notice?

3. What is the estimate from this sample of  $\mathbb{P}(\text{Diabetic}|\text{Postive test})$ ?

*Click Flip to see this*

4. Was this sample unusual?

*Select sample 10 times to check. What do you notice?*

Why is  $\mathbb{P}(\text{Diabetic}|\text{Postive test})$  so small?

5. Some sub-populations have a higher prevalence of diabetes.

*Clear your results and change the prevalence from 6% to 12%.*

Does  $\mathbb{P}(\text{Diabetic}|\text{Postive test})$  change? How?

6. If you were being tested for diabetes what would you like  $\mathbb{P}(\text{Diabetic}|\text{Positive test})$  to be?

*Try adjusting the prevalence.*

What prevalence would be needed for this test to be correct with the probability you desire?

7. Other tests have a higher sensitivity and specificity.

*Change both the sensitivity and specificity to 99%.*

What effect does this have on the prevalence needed for you to get your desired  $\mathbb{P}(\text{Diabetic}|\text{Positive test})$ ?

Is it still possible for  $\mathbb{P}(\text{Diabetic}|\text{Positive test})$  to be small?

8. Is it ever possible for  $\mathbb{P}(A|B) = \mathbb{P}(B|A)$ ? Try it!

9. Refer back to the questions in the box at the top of the first page ... Summarise what you have learnt ☺