**Prosecutor’s Fallacy and Mammograms Case Study**

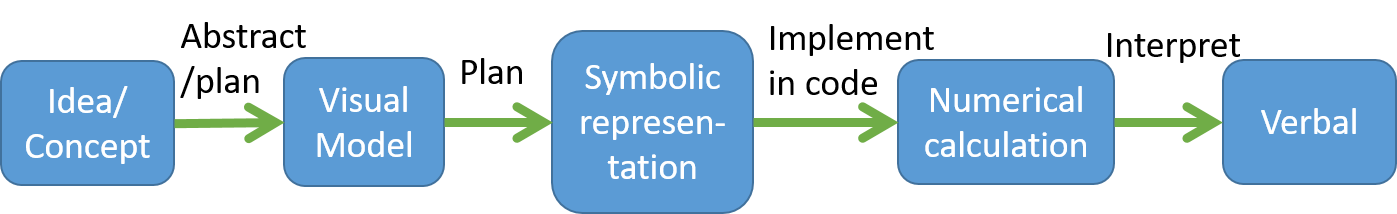
An introduction to Prosecutor’s Fallacy, and False Positives/Negatives. by Dr. Diaz Eaton, Bates College, for use in DCS 105A, and inspired by [Calling Bull](https://callingbullshit.org/) & [Calculated Risks](https://www.amazon.com/Calculated-Risks-Know-Numbers-Deceive/dp/0743254236).

Learning Objectives:

* Introduction to probability concepts such as Bayes Theorem, conditional probabilities, false positives and false negatives.
* Explore how representation in the data set can affect interpretation of p-values.
* Practice automating repetitive calculations through spreadsheets (and optional challenge - through R syntax and RStudio).

**MODELING/PROGRAMMING**

Recall that **modeling and programming** are broad terms that encompasses a process including elements such as planning, implementation, testing, revision, feedback, etc., for some purpose. In the case we are considering below, our process of modeling/programming is:



**CONTEXT**

In the Calling Bull videos, we heard that p-values only tell us a part of the information, and it is easy to be mislead by looking at p-value alone. This could be done intentionally - but is often a misunderstanding of what a DNA match or a p-value tells us. The **p-value** tells us how likely it is that this is a result of chance alone (sampling from the null distribution - the distribution that the null hypothesis and the statistical test would assume).

So what does it mean if a DNA match comes up positive in a database? Image matching in a database? Should make drug testing for all mandatory? Drug testing for foodstamp recipients? Drug testing for senators? What about mammograms for all?

**Probability**

Suppose we consider some **variable** that represents the **state** of a coin flip. There are two possible **outcomes**: “heads” or “tails.” The set of {“heads, “tails”} is sometimes referred to as the **sample space**. The **probability** of a particular event, A, occurring can be written as ***P(A),*** which is a nice shortcut. If all outcomes are equally likely, P(A) is calculated as the number of times A is in the sample space/number of outcomes in the total sample space.

**Independence of Events**

When we have two probabilities that are **independent**, P(B happens and C happens ) = P(B happens)\*P(C happens). For example, when we throw a 6-sided die and flip coin, the probability that you roll a 6 and flip a “heads” are are not related or contingent upon each other. Therefore P(6 and heads) = P(die is 6)\*P(coin is heads) = ⅙ \* ½ = 1/12. Also, the probability of observing a heads is always ½ regardless of what the die shows, so P(coin is heads) = ½.

In the case of Prosecutor’s Fallacy is that you have two **variables** under consideration - the state of the accused and the state of the DNA test. Each variable has two possible **outcomes** - guilt or innocence of the accused, and a match or non-match of the DNA test - that are ***not*** **independent**. In the case of events that are not independent, we might be interested in the probability an event occurs on its own, or the probability of an event given a particular state of the related event.

For example, suppose we are thinking about the weather and Batsie umbrella usage. These events are clearly related and not independent. The two events are daily weather (states: rain or no rain) and Batsie umbrella usage (states: umbrella or no umbrella).

**\*Coding professional practice - name the parameters close to what they mean. Use # (aka “comment out”) before descriptors that explain what you are doing and/or the units, or where the figure comes from.**

If we want to ponder the probability of using an umbrella, there are a few ways that we could think about it:

P(umbrella) = probability of umbrella usage on any given day regardless of weather

P(umbrella | rain) = probability of seeing an umbrella “given” a rainy day

P(umbrella | no rain) = probability of seeing an umbrella “given” a day with no rain

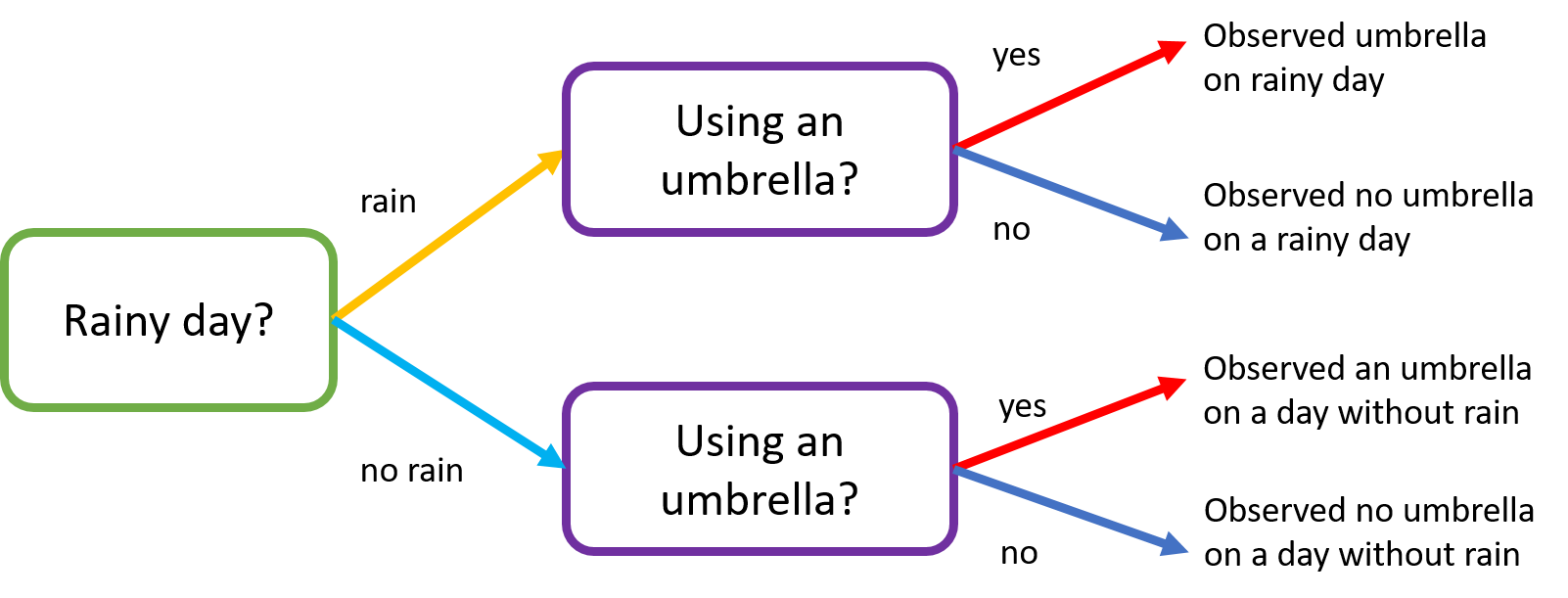
Each of the probabilities above have different interpretations and different values. WARNING - BE CAREFUL! This is the source of the **Prosecutor's Fallacy**.

\*Also, please note that | symbol means something different here than in R\*

**Visualizing probabilities**

There are a couple ways to organize all of the information about probabilities.

**Option 1 - Tree Diagram:**



P(umbrella) = (sum of people that fit into red arrow categories)/(total people observed)

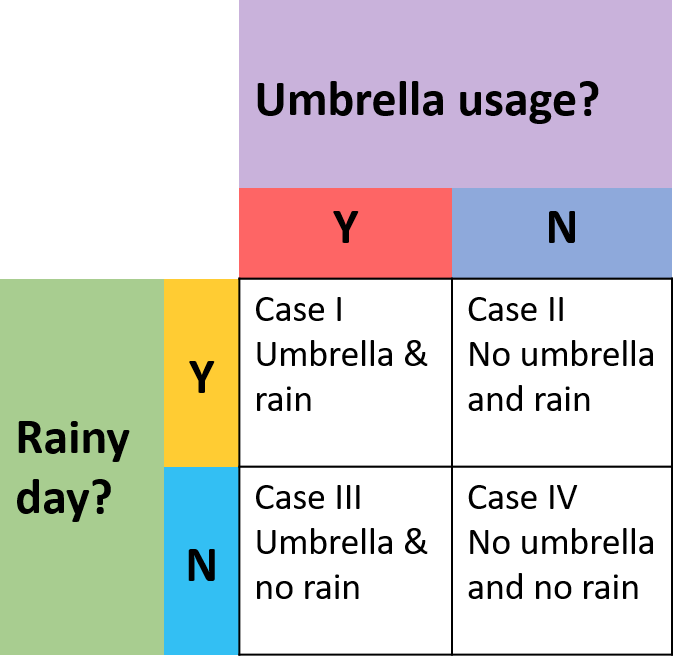
P(umbrella | rain) = (# people in top red arrow category)/(total rainy days)

P(umbrella | no rain) = (# people in bottom red arrow category)/(total non-rainy days)

**Option 2 - Box Diagram:**

This box has been color coded to match option 1.

P(umbrella) = (Case I + Case III)/total observations

P(umbrella | rain) = #Case I/(#yellow: I + II)

P(umbrella | no rain) =#Case III/(#cyan: III + IV)

You may also recognize this from the Calling Bull Videos. To translate with the videos:

Ho null hypothesis - presumed innocence

Purple umbrella - experimental result - DNA match

* Red (match)
* Indigo (non match)

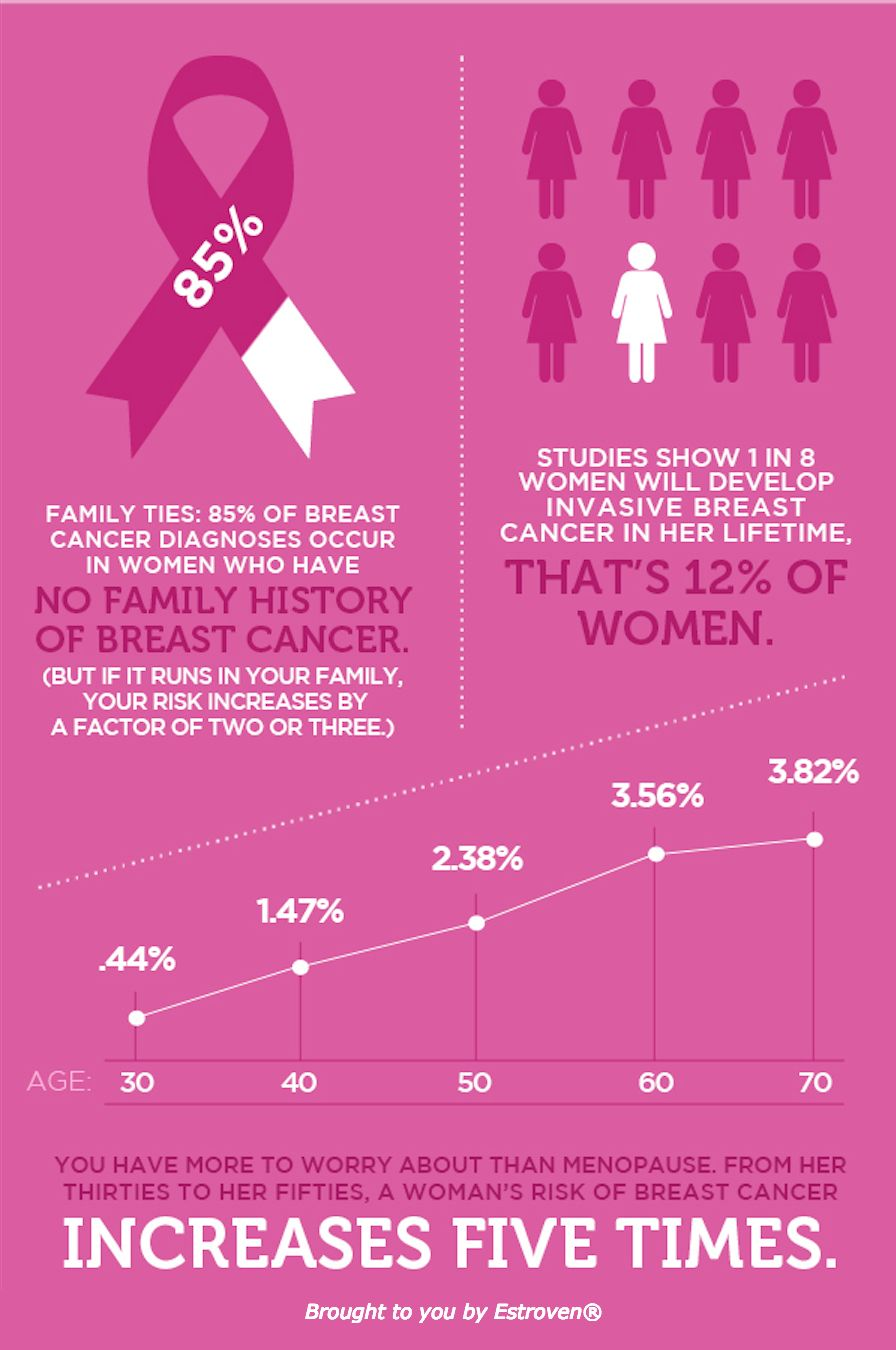
Green rainy day - actual reality - guilt of the accused

* Yellow (guilty)
* Cyan (innocent)

In this particular visualization, it is easy to spot false positives and false negatives.

A **false negative** is Case II & a **false positive** is Case III.

**Discuss**

Confirm the following probability interpretations and calculations (do these formulas make sense?): 

P(match) = (Case I + Case III)/total observations

P(innocent) = (Case III + Case IV)/total observations

P(match | innocent) = Case III/(Case III +Case IV)

\*Note that this is also the p-value explained in the video and [this link](https://stats.stackexchange.com/questions/105757/is-p-value-also-the-false-discovery-rate).

Any other probabilities?

**Mammogram screenings for all?**

Here’s a meme on social media. This looks like serious business. But we also know from prior readings that healthcare is full of bullshit. Let’s think about this meme for a bit.

**Discuss**

In your groups, talk about how this makes you feel, look at the arguments, the data visualization techniques. What did you learn (assuming the information here is true). What are some questions you have? What are some criticisms you might have?

**YOUR BS-ometer MISSION**

Let’s think about the benefits of getting a mammogram based on what we know from numbers. There was a big study in Germany - 3-4 million women get a mammogram each year and we have the data. Of the 4 million women, approximately 34,000 have breast cancer that is detected, 1790 have breast cancer that is not detected, 300,000 do not have breast cancer but is read as positive on a mammogram, and the rest do not have breast cancer and the mammogram is also negative (from Calculated Risks, by Gigerenzer).

Your task will be to create a program that will automatically calculates each probability and conditional probabilities from this given set of observations (the counts at the tips of the decision tree or in the boxes). You will then interpret what these probabilities might mean for policy and to compare that to the accuracy/feelings conveyed by the meme above.

**PLAN**

\***Programming professional practice - intentionally plan your code before you start writing. This can be done through visual sketches, thinking about what inputs, variables, and parameters, you will need to get the output of interest, and/or writing psuedocode.**

In this case,

1. Create a visualization - probably the box works better in this case - as a way to think about what your counts mean.
2. Figure out what probabilities and calculations/formulas you will need to implement.
3. Code this in a spreadsheet or (Optional) code this in RStudio.
4. Interpret.

**IMPLEMENT**

(below are implementation guidelines for spreadsheets only - coding this in RStudio is an optional challenge - see below)

**Prepare:**

Start a new GoogleSheet and share it with teammates.

Add a title and other information to the top

In cells at the top put:

DCS 105 A Calling Bull, Feb 6, 2019

Mammograms Case Study

Team members names

**Step 1:**

#Create a visualization - probably the box works better in this case - as a way to think about what your counts mean. Feel free to use a computer or to use a sketch pad - but draw it up on a board next to you once you have it.

**Step 2:**

Figure out what probabilities and calculations/formulas you will need to implement.

**\*Coding professional practice - build in ways to check your code to make sure it is functioning as intended.**

One formula you can use to check is Bayes Theorem for conditional probabilities: .

**Step 3:**

#Code this in a spreadsheet

Think about mirroring the box diagram in your spreadsheet cells. Make sure you’ve checked your work in some way.

**Step 4:**

#Interpret

Examine the output, record observations, discuss as a team.

**ANALYZE**

Compare and contrast this to the meme you saw above. Any surprises? Can you find any memes that portray some of your additional observations?

**MODIFY/PLAY**

1. **Discuss & Plan:** How would you modify the code you wrote if there were less false positives (be careful not to change the total # of individuals). What if breast cancer were more common?
2. **Implement**
3. **Analyze output:** 
   1. What answer did you get?
   2. What does it mean?

**CHALLENGE (Optional)**

✨For those of you who have finished the DataCamp tutorial on functions, feel free to try implementing the same kind of program in R. In other words, code a function,

function\_name(caseI, caseII, caseIII, caseIV), that will do the same as the spreadsheet exercise.

**ADDITIONAL Reading and resources**

Mammograms

<https://www.healthnewsreview.org/2017/01/one-key-detail-youre-unlikely-to-see-in-news-stories-about-mammography-screening-guidelines/>

Sensitivity vs Specificity

<https://en.wikipedia.org/wiki/Sensitivity_and_specificity>

Calculated Risks by Greg Gigerenzer

<https://www.amazon.com/Calculated-Risks-Know-Numbers-Deceive/dp/0743254236>