Chapter 3

Probability - the nuts and bolts of Bayesian inference

3.1 Messy probability density

Suppose that a probability density is given by the following function:

$$f(X) = \begin{cases} 1, & \text{if } 0 \leq X < 0.5 \\ 0.2, & \text{if } 0.5 \leq X < 1 \\ 0.8(X-1), & \text{if } 1 \leq X < 2 \\ 0, \text{otherwise} & \end{cases}$$

Problem 3.1.1. Demonstrate that the above density is a valid probability distribution.

This can be done by summing the areas from each of the three parts,

$$\text{area} = 1 \times 0.5 + 0.2 \times 0.5 + 0.5 \times 1 \times 0.8 = 1 \tag{3.1}$$

Problem 3.1.2. What is the probability that $0.2 \leq X \leq 0.5$?

The relevant area is,

$$Pr(0.2 \leq X \leq 0.5) = 0.3 \tag{3.2}$$

Problem 3.1.3. Find the mean of the distribution.

To do this we need to integrate,

$$\mathbb{E}(X) = \int_0^2 x f(x) \mathrm{d}x \tag{3.3}$$
$$= \int_0^{0.5} x \mathrm{d}x + \int_{0.5}^1 0.2x \mathrm{d}x + \int_1^2 (x-1)0.8x \mathrm{d}x \tag{3.4}$$
$$= [0.5x^2]_0^{0.5} + [0.1x^2]_{0.5}^1 + 0.8[\frac{1}{3}x^3 - \frac{1}{2}x^2]_1^2 \tag{3.5}$$
$$= 0.5 \times 0.5^2 + 0.1(1 - 0.5^2) + 0.8(\frac{8}{3} - 2 - \frac{1}{3} + \frac{1}{2}) \tag{3.6}$$
$$\approx 0.867 \tag{3.7}$$

Problem 3.1.4. What is the median of the distribution?

This is the point at which $Pr(X < a) = 0.5$, which happens to occur at $a = 0.5$.

3.2 Keeping it discrete

Suppose that the number of heads obtained X in a series of N coin flips is described by a binomial distribution:

$$Pr(X = K|\theta) = \binom{N}{K} \theta^K (1-\theta)^{N-K}, \tag{3.8}$$

where $\binom{N}{K} = \frac{N!}{K!(N-K)!}$ is the binomial coefficient and θ is the probability of obtaining a heads on any particular throw.

Problem 3.2.1. Suppose that $\theta = 0.5$ (that is, the coin is fair). Calculate the probability of obtaining 5 heads in 10 throws.

This is given by,

$$Pr(X = 5|\theta = 0.5) = \binom{10}{5} 0.5^5 0.5^5 \approx 0.246 \tag{3.9}$$

Problem 3.2.2. Calculate the probability of obtaining fewer than 3 heads.

This is given by,

$$Pr(X < 3|\theta = 0.5) = Pr(X = 0|\theta = 0.5) + Pr(X = 1|\theta = 0.5) + Pr(X = 2|\theta = 0.5) \tag{3.10}$$
$$= 0.5^{10} + 10 \times 0.5^{10} + 45 \times 0.5^{10} \tag{3.11}$$
$$\approx 0.055 \tag{3.12}$$

Problem 3.2.3. Find the mean of this distribution. (You can either derive the mean of this distribution or take it as given that $\mathbb{E}(X) = N\theta$.)

The mean here is given by $\mathbb{E}(X) = N\theta = 10 \times 0.5 = 5$.

Problem 3.2.4. Suppose I flip another coin with $\theta = 0.2$. What is the probability that I get more than 8 heads?

$$\begin{aligned} Pr(X_2 > 8|\theta = 0.2) &= Pr(X = 9|\theta = 0.2) + Pr(X = 10|\theta = 0.2) && (3.13)\\ &= 10 \times 0.2^9 0.8^1 + 0.2^{10} && (3.14)\\ &\approx 4 \times 10^{-6} && (3.15) \end{aligned}$$

Problem 3.2.5. What is the probability that I obtain fewer than 3 heads in 10 flips of the first coin, and more than 8 heads with the second?

The two events are independent and so,

$$\begin{aligned} Pr(X_1 < 3, X_2 > 8|\theta_1 = 0.5, \theta_2 = 0.2) &= Pr(X_1 < 3|\theta_1 = 0.5) \times Pr(X_2 > 8|\theta_2 = 0.2) && (3.16)\\ &\approx 0.055 \times 4 \times 10^{-6} && (3.17)\\ &\approx 2 \times 10^{-7} && (3.18) \end{aligned}$$

3.3 Continuously confusing

Suppose that the time that elapses before a particular component on the Space Shuttle fails can be modelled as being exponentially distributed:

$$p(t|\lambda) = \lambda e^{-\lambda t}, \tag{3.19}$$

where $\lambda > 0$ is a rate parameter.

Problem 3.3.1. Show that the above distribution is a valid probability density.

To do this we integrate,

$$\begin{aligned} \int_0^\infty \lambda e^{-\lambda t} \mathrm{d}t &= [-e^{-\lambda t}]_0^\infty && (3.20)\\ &= 1 && (3.21) \end{aligned}$$

Problem 3.3.2. Find the mean of this distribution.

To do this we integrate (using integration by parts),

$$\int_0^\infty \lambda t e^{-\lambda t} \mathrm{d}t = \frac{1}{\lambda} \tag{3.22}$$

Problem 3.3.3. Suppose that $\lambda = 0.2$ per hour. Find the probability that the component fails in the first hour of flight.

To do this we integrate,

$$\begin{aligned} Pr(0 \leq t \leq 1 | \lambda = 0.2) &= [-e^{-0.2t}]_0^1 && (3.23)\\ &= 1 - e^{-0.2} && (3.24)\\ &\approx 0.18 && (3.25) \end{aligned}$$

Problem 3.3.4. What is the probability that the component survives for the first hour but fails during the second?

To do this we integrate,

$$\begin{aligned} Pr(1 \leq t \leq 2 | \lambda = 0.2) &= [-e^{-0.2t}]_1^2 && (3.26)\\ &= e^{-0.2} - e^{-0.4} && (3.27)\\ &\approx 0.148 && (3.28) \end{aligned}$$

Problem 3.3.5. What is the probability that the component fails during the second hour given that it has survived the first?

This is a conditional probability,

$$\begin{aligned} Pr(1 \leq t \leq 2 | t \geq 1, \lambda = 0.2) &= \frac{Pr(1 \leq t \leq 2 | \lambda = 0.2)}{Pr(t \geq 1 | \lambda = 0.2)} && (3.29)\\ &= \frac{0.148}{0.82} && (3.30)\\ &\approx 0.18 && (3.31) \end{aligned}$$

We could have obtained this from the memoryless property of the exponential distribution (see next question).

Problem 3.3.6. Show that the probability of the component failing during the $(n+1)$th hour given that it has survived n hours is always 0.18.

This is a conditional probability,

$$\begin{aligned} Pr(n \leq t \leq n+1 | t \geq n, \lambda = 0.2) &= \frac{Pr(n \leq t \leq n+1 | \lambda = 0.2)}{Pr(t \geq n | \lambda = 0.2)} && (3.32) \\ &= \frac{[-e^{-0.2t}]_n^{n+1}}{1 - [-e^{-0.2t}]_0^n} && (3.33) \\ &= \frac{e^{-0.2n} - e^{-0.2(n+1)}}{e^{-0.2n}} && (3.34) \\ &= 1 - e^{-0.2} && (3.35) \\ &\approx 0.18, && (3.36) \end{aligned}$$

and so we have demonstrated the memoryless property of the exponential distribution.

3.4 The boy or girl paradox

The boy or girl paradox was first introduced by Martin Gardner in 1959. Suppose we are told the following information:

Problem 3.4.1. Mr Bayes has two children. The older child is a girl. What is the probability that both children are girls?

There are two potentialities that include the older child being a girl: boy-girl or girl-girl. Hence the probability that both children are girls is $\frac{1}{2}$.

Problem 3.4.2. Mr Laplace has two children. At least one of the children is a girl. What is the probability that both children are girls?

Here there are three potentialities: boy-girl, girl-boy or girl-girl. This means that there is a $\frac{1}{3}$ probability that both children are girls. Note that there has been some controversy over the asking of this question. See `https://en.wikipedia.org/wiki/Boy_or_Girl_paradox`.

3.5 Planet Scrabble

On a far-away planet suppose that people's names are always two letters long, with each of these letters coming from the 26 letters of the Latin alphabet. Suppose that there are no constraints on individuals' names, so they can be composed of two identical letters, and there is no need to include a consonant or a vowel.

Problem 3.5.1. How many people would need to be gathered in one place for there to be a 50% probability that at least two of them share the same name?

There are $26 \times 26 = 676$ possible names out there. Let's start with four people, and work out the probability that two of them do *not* share the same name, i.e. $X = 0$, where X is the number of occurrences of people with shared names in the group.

$$Pr(X = 0) = \frac{676}{676} \times \frac{675}{676} \times \frac{674}{676} \times \frac{673}{676} \tag{3.37}$$

$$= \left(\frac{1}{676}\right)^4 \times (676 \times 675 \times 674 \times 673) \tag{3.38}$$

$$= \left(\frac{1}{676}\right)^4 \times \frac{676!}{672!} \tag{3.39}$$

Or more generally we have that for n individuals,

$$Pr(X = 0) = \frac{{}_{676}\mathcal{P}_n}{676^n}, \tag{3.40}$$

where ${}_{676}\mathcal{P}_n = \frac{676!}{(676-n)!}$ is the permutation coefficient. The graph of this function is shown in Figure 3.1, where we see that if 31 or more people are gathered in a room there is a probability that exceeds 50% that two will share the same name.

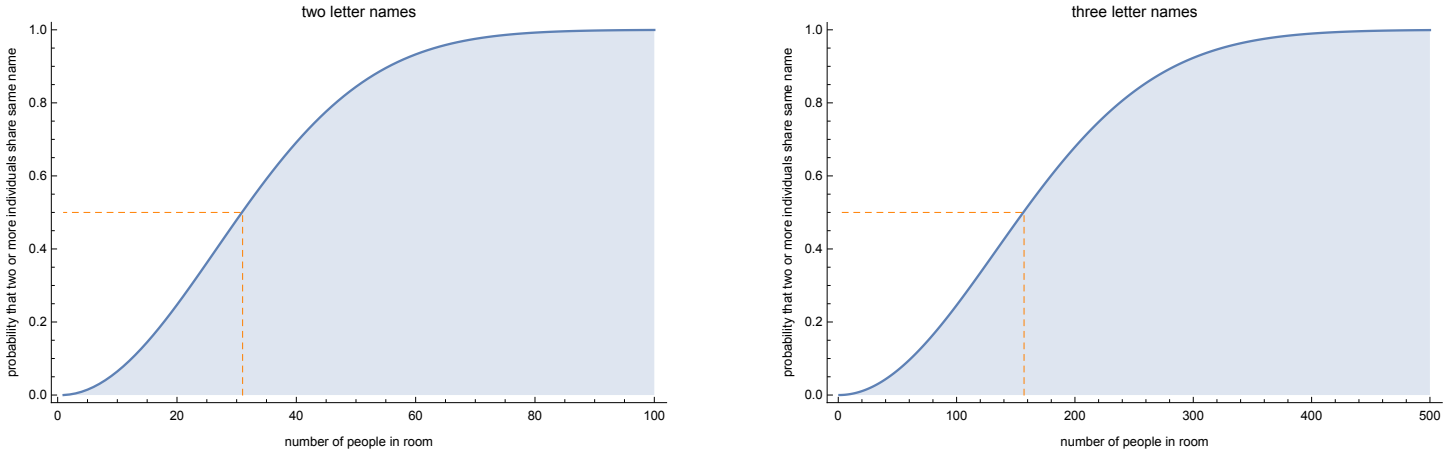

Figure 3.1: The probability that two or more people share the same name in a two- (left) and three- (right) letter name world.

Problem 3.5.2. Suppose instead that the names are composed of three letters. Now how many people would need to be gathered in one place for there to be a 50% probability that at least two of them share the same name?

The same analysis as above applies although now with $26 \times 26 \times 26 = 17,576$ possible names out there. In this case we find 157 is the minimum number of people required for there to be at least a 50% probability of two sharing the same name (see Figure 3.1).

3.6 Game theory

A game show presents contestants with four doors: behind one of the doors is a car worth \$1000; behind another is a forfeit whereby the contestant must pay \$1000 out of their winnings thus far on the show. Behind the other two doors there is nothing. The game is played as follows:

1. The contestant chooses one of four doors.
2. The game show host opens another door, always to reveal that there is nothing behind it.
3. The contestant is given the option of changing their choice to one of the two remaining unopened doors.
4. The contestant's final choice of door is opened, to their delight (a car!), dismay (a penalty), or indifference (nothing).

Assuming that:

- the contestant wants to maximise their expected wealth, and
- the contestant is risk-averse,

what is the optimal strategy for the contestant?

This question hinges on deriving the distribution of outcomes under either remaining, or changing, after the game show host has opened the empty door. This can be answered through application of Bayes' rule, but I prefer here to describe more intuitively what is happening.

Imagine we are considering repeating the show a number of times. There are three possibilities for the initial choice of door:

- $\frac{1}{4}$ of the time, the door hides the car.
- $\frac{1}{2}$ of the time, the door hides a null.
- $\frac{1}{4}$ of the time, the door hides the penalty.

Considering now each of these in turn:

If the door contains the car, then the other three doors are two nulls, and the penalty. The game show host opens one of the nulls, meaning that only one null and the penalty remain. In this circumstance, if you stay put, you definitely obtain the car. If you change, you get a null with probability $\frac{1}{2}$ and similarly for the penalty.

If the door contains a null, then the other three doors are one null, one penalty and one car. When the host opens the remaining null, then the other two doors are the car, and the penalty. This is the key step. By remaining, you gain/lose nothing, whereas if you change you face risk; you get the car with probability $\frac{1}{2}$ and similarly for the penalty. Both of these choices have the same expected payoff, but the latter increases risk.

Finally, if the door contains the penalty, then the other three doors are two nulls, and the car. The game show host opens one of the nulls, meaning that only one null and the car remain. In this circumstance, if you stay put, you definitely obtain the penalty. If you change, you get a null with probability $\frac{1}{2}$ and similarly for the car.

We can now write down probability distributions for the outcomes given each possible action. For remaining, the probabilities are what they were if you hadn't received any new information, in other words $(p(car), p(null), p(penalty)) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. Whereas if you change:

$$p(car) = \frac{1}{4} \times 0 + \frac{1}{2} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} = \frac{3}{8} \tag{3.41}$$

$$p(null) = \frac{1}{4}\frac{1}{2} + \frac{1}{2} \times 0 + \frac{1}{4} \times \frac{1}{2} = \frac{2}{8} \tag{3.42}$$

$$p(penalty) = \frac{1}{4} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} + \frac{1}{4} \times 0 = \frac{3}{8} \tag{3.43}$$

So in summary, we get face $(p(car), p(null), p(penalty)) = (\frac{3}{8}, \frac{2}{8}, \frac{3}{8})$ by changing.

Both of these two outcomes face the same expected return of \$0:

$$\mathbb{E}[return|stay] = \frac{1}{4} \times \$1,000 + \frac{1}{2} \times \$0 + \frac{1}{4} \times -\$1,000 \tag{3.44}$$

$$= \$0 \tag{3.45}$$

$$\mathbb{E}[return|change] = \frac{3}{8} \times \$1,000 + \frac{2}{8} \times \$0 + \frac{3}{8} \times -\$1,000 \tag{3.46}$$

$$= \$0 \tag{3.47}$$

However, what is different is the variance in return from each decision. The variance in return is greater by changing than it is by staying, since the latter places more weight on the risky outcomes. Obviously, these can be calculated explicitly using the same methodology as above.

If the individual is risk-averse, then he prefers the less risky outcome of *remaining*, given that they both have the same return.

3.7 Blood doping in cyclists

Suppose, as a benign omniscient observer, we tally up the historical cases where professional cyclists either used or did not use blood doping, and either won or lost a particular race. This results in the probability distribution shown in Table 3.1.

Problem 3.7.1. What is the probability that a professional cyclist wins a race?

This is the marginal given by: $p(won) = p(won, dope) + p(won, clean) = 0.05 + 0.1 = 0.15$.

	Lost	**Won**
Clean	0.70	0.05
Doping	0.15	0.10

Table 3.1: The historical probabilities of behaviour and outcome for professional cyclists.

Problem 3.7.2. What is the probability that a cyclist wins a race, given that they have cheated?

$$p(won|doped) = \frac{p(doped, won)}{p(doped)} \tag{3.48}$$

$$= \frac{0.1}{0.25} = 0.4 \tag{3.49}$$

Problem 3.7.3. What is the probability that a cyclist is cheating, given that they win?

$$p(doped|won) = \frac{p(won|doped)p(doped)}{p(won)} \tag{3.50}$$

$$= \frac{0.4 \times 0.25}{0.15} \tag{3.51}$$

$$= \frac{2}{3} \tag{3.52}$$

Now suppose that drug testing officials have a test that can accurately identify a blood-doper 90% of the time. However, it incorrectly indicates a positive for clean athletes 5% of the time.

Problem 3.7.4. If the officials care only about the proportion of people correctly identified as dopers, should they test all the athletes or only the winners?

Here we want to compare $p(doped|positive, group)$ across $group \in \{everyone, winners\}$. For everyone, this is simple and given by:

$$p(doped|positive) = \frac{p(positive|doped)p(doped)}{p(positive)} \tag{3.53}$$

$$= \frac{p(positive|doped)p(doped)}{p(positive, doped) + p(positive, clean)} \tag{3.54}$$

$$= \frac{p(positive|doped)p(doped)}{p(positive|doped)p(doped) + p(positive|clean)p(clean)} \tag{3.55}$$

$$= \frac{0.9 \times 0.25}{0.9 \times 0.25 + 0.05 \times 0.75} \tag{3.56}$$

$$\approx 0.86 \tag{3.57}$$

Whereas for the winners:

$$p(doped|positive, won) = \frac{p(doped, positive|won)}{p(positive|won)} \tag{3.58}$$

$$\tag{3.59}$$

We will proceed to calculate each of these bits in turn. Via Bayes' rule:

$$\begin{aligned} p(doped, positive|won) &= \frac{p(won|doped, positive)p(doped, positive)}{p(won)} && (3.60)\\ &= \frac{p(won|doped)p(doped, positive)}{p(won)} && (3.61)\\ &= \frac{p(won|doped)p(positive|doped)p(doped)}{p(won)} && (3.62)\\ &= \frac{0.4 \times 0.9 \times 0.25}{0.15} && (3.63)\\ &= 0.6 && (3.64) \end{aligned}$$

We have got the second line from the first by assuming that there is a conditional independence between winning and testing positive, once we account for their drug status. This is a fairly safe assumption, unless of course winners are more effective at hiding their drug use!

Now for the last bit:

$$\begin{aligned} p(positive|won) &= p(positive, doped|won) + p(positive, clean|won) && (3.65)\\ &= p(positive|doped, won)p(doped|won) + p(positive|clean, won)p(clean|won) && (3.66)\\ &= p(positive|doped)p(doped|won) + p(positive|clean)p(clean|won) && (3.67)\\ &= 0.9 \times \frac{2}{3} + 0.05 \times \frac{1}{3} && (3.68)\\ &\approx 0.62 && (3.69) \end{aligned}$$

Combining these two, we have $p(doped|positive, won) = \frac{0.6}{0.62} \approx 0.97$. Hence we should only test the winners. This makes intuitive sense, since they are a group which have a higher than average percentage of dopers.

Problem 3.7.5. If the officials care five times as much about the number of people who are falsely identified as they do about the number of people who are correctly identified as dopers, should they test all the athletes or only the winners?

Now we need to specify a utility function of the form:

$$\begin{aligned} U(group) &= n(doped|positive, group) - 5n(clean|positive, group) && (3.70)\\ &= n(group)\left[p(doped|positive, group) - 5p(clean|positive, group)\right] && (3.71)\\ &= n(group)\left[p(doped|positive, group) - 5(1 - p(doped|positive, group))\right] && (3.72)\\ &= n(group)\left[6p(doped|positive, group) - 5\right] && (3.73)\end{aligned}$$

Calculating this for everyone, we have:

$$\begin{aligned} U(total) &= n(total)\left[6p(doped|positive) - 5\right] && (3.74)\\ &\approx n(total)\left[6 \times 0.86 - 5\right] && (3.75)\\ &= 0.16n(total) && (3.76)\end{aligned}$$

For only the winners' group, we have:

$$\begin{aligned} U(won) &= n(total) \times p(won)\left[6p(doped|positive, won) - 5\right] && (3.77)\\ &\approx n(total) \times 0.15\left[6 \times 0.97 - 5\right] && (3.78)\\ &\approx 0.12n(total) && (3.79)\end{aligned}$$

So in this case they should test everyone.

Problem 3.7.6. What factor would make the officials choose the other group? (By factor, we mean the number 5 in the previous problem.)

We can calculate $U(g)$ of a group g as a function of α; the factor:

$$U(g) = n(g)\left[(1 + \alpha)p(D|+, g) - \alpha\right] \tag{3.80}$$

This means we can calculate:

$$\begin{aligned} U(total) &= n(total)\left[(1 + \alpha)p(D|+) - \alpha\right]\\ &= n(total)\left[(1 + \alpha)0.86 - \alpha\right]\\ &= n(total)\left[0.86 - 0.14\alpha\right]\end{aligned}$$

And:

$$\begin{aligned} U(win) &= n(win)\left[(1 + \alpha)p(D|+, W) - \alpha\right]\\ &= n(win)\left[(1 + \alpha)0.97 - \alpha\right]\\ &= 0.15n(total)\left[0.97 - 0.03\alpha\right]\end{aligned}$$

Now finding α such that $U(win) > U(total)$ we find, $\alpha > 5.23$.

3.8 Breast cancer revisited

Suppose that the prevalence of breast cancer for a randomly chosen 40-year-old woman in the UK population is about 1%. Further suppose that mammography has a relatively high sensitivity to breast cancer, where in 90% of cases the test shows a positive result if the individual has the disease. However, the test also has a rate of false positives of 8%.

Problem 3.8.1. Show that the probability that a woman tests positive is about 9%.

Here we need to determine $p(+)$, which is the marginal probability of testing positive. We get this by marginalising the joint probability, $p(+, C)$,

$$p(+) = \underbrace{p(+|C) \times p(C)}_{\text{positive and cancer}} + \underbrace{p(+|NC) \times p(NC)}_{\text{false positive}} \tag{3.81}$$

$$= 0.9 \times 0.01 + 0.08 \times 0.99 \tag{3.82}$$

$$= 8.8\% \tag{3.83}$$

Problem 3.8.2. A woman tests positive for breast cancer. What is the probability she has the disease?

Use Bayes' rule to find this quantity,

$$p(C|+) = \frac{p(+|C) \times p(C)}{p(+)} \tag{3.84}$$

$$= \frac{0.9 \times 0.01}{0.088} \tag{3.85}$$

$$\approx 10\% \tag{3.86}$$

Problem 3.8.3. Draw a graph of the probability of having a disease, given a positive test, as a function of (a) the test sensitivity (true positive rate) (b) the false positive rate, and (c) the disease prevalence. Draw graphs (a) and (b) for a rare (1% prevalence) and a common (10% prevalence) disease. What do these graphs imply about the relative importance of the various characteristics of medical tests?

We can write down the probability as we did before using Bayes' rule,

$$p(C|+) = \frac{p(+|C) \times p(C)}{p(+|C) \times p(C) + p(+|NC) \times (1 - p(C))} \tag{3.87}$$

which we can then graph across the three variables in this expression (fig. 3.2). For rare diseases this illustrates the weak dependence of the test on the true positive rate, and the strong dependence

on the false positive rate. So for diseases that are rare, the most important thing is ensuring that the test has a low false positive rate. This makes intuitive sense, because for rare diseases the number of false positives quickly dwarfs the true positives. Whereas for more common diseases this factor is not so important; ensuring that the test sensitivity is high is a more pressing concern.

Of course the previous analysis does not give a utility to each outcome, so it is hard to make conclusions about the optimality of the test; Bayesian decision theory would be needed here to determine optimal testing parameters!

The plot of probability of testing positive for a disease versus disease prevalences illustrates that this medical test conveys most information for common diseases (15-30% prevalence). This is because there is the biggest gain in information between not having the test (black line) and post-test information (blue line).

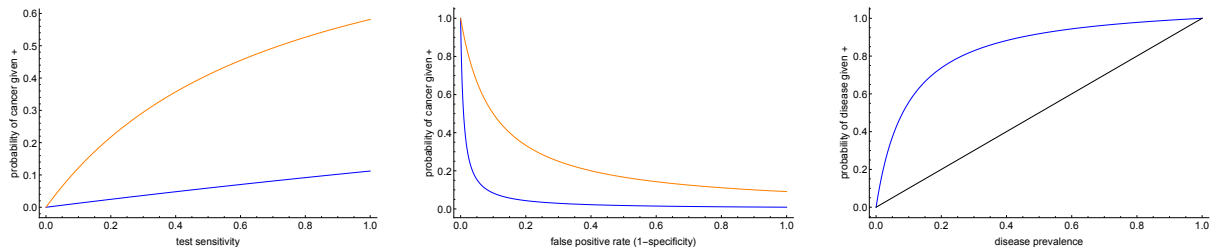

Figure 3.2: A plot of the probability of disease given a positive test result as a function of a. test sensitivity, b. the false positive rate for rare (blue) and common (orange) diseases, and c. the disease prevalence. For rare diseases we assume prevalence is 1%, and for common ones we assume 10% prevalence. For the right hand plot we assume a sensitivity of 90% and a specificity of 92%.

Problem 3.8.4. Assume the result of a mammography is independent when retesting an individual (probably a terrible assumption!). How many tests (assume a positive result in each) would need to be undertaken to ensure that the individual has a 99% probability that they have cancer?

Let's start by working out the probability of cancer for two positive test results,

$$p(C|++) = \frac{p(++|C) \times p(C)}{p(++)} \tag{3.88}$$

$$= \frac{p(+|C) \times p(+|C) \times p(C)}{p(++)} \tag{3.89}$$

where the marginal probability of two positive test results is similar to before,

$$p(++) = p(+|C) \times p(+|C) \times p(C) + p(+|NC) \times p(+|NC) \times p(NC) \tag{3.90}$$

$$= 0.9^2 \times 0.01 + 0.08^2 \times 0.99 \tag{3.91}$$

$$\approx 0.0144 \tag{3.92}$$

and so $p(C|++) \approx \frac{0.9^2 \times 0.01}{0.0144} = 56\%$. Using the above formulae we note that the probability for the case of n tests is given by,

$$p(C|+^n) = \frac{p(+^n|C) \times p(C)}{p(+^n)} \tag{3.93}$$

$$= \frac{p(+|C)^n \times p(C)}{p(+|C)^n \times p(C) + p(+|NC)^n \times p(NC)} \tag{3.94}$$

$$= \frac{0.9^n \times 0.01}{0.9^n \times 0.01 + 0.08^n \times 0.99} \tag{3.95}$$

If we graph this function we find that it is logistic-sigmoid shaped (Figure 3.3), and that after four tests we have reached the required threshold.

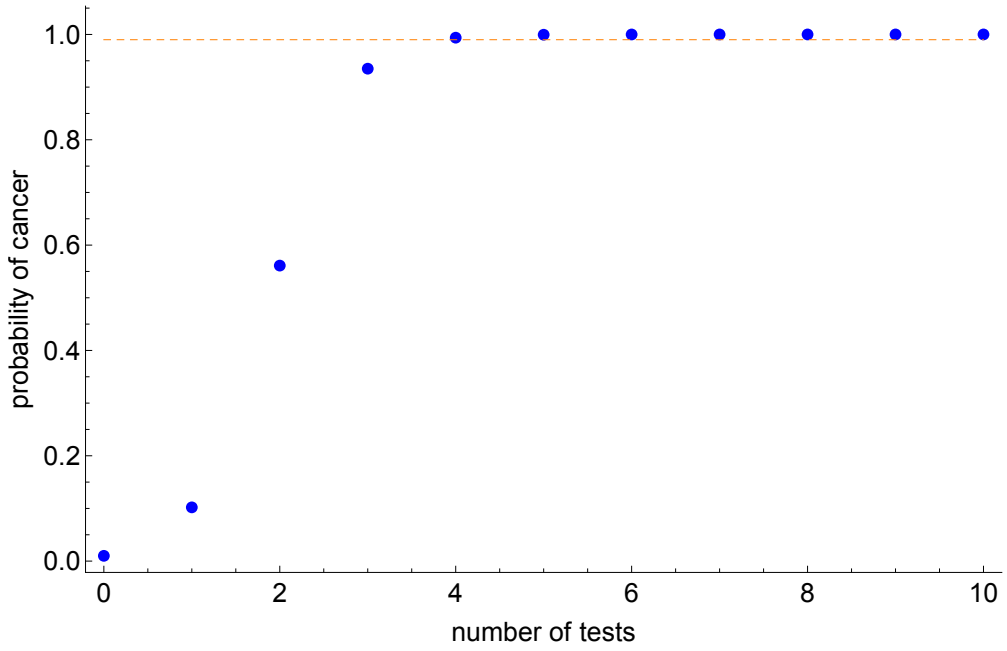

Figure 3.3: A plot of the probability of cancer as a function of the number of tests.

Problem 3.8.5. Now we make the more realistic assumption that the probability of testing positive in the nth trial depends on whether positive tests were achieved in the $(n+1)$th trials, for both individuals with cancer and those without. For a cancer status $\kappa \in \{C, NC\}$:

$$p(n+|(n-1)+, \kappa) = 1 - (1 - p(+|\kappa))e^{-(n-1)\epsilon} \tag{3.96}$$

where $n+$ denotes testing positive in the nth trial, $p(+|\kappa)$ and $\epsilon \geq 0$ determine the persistence in test results. Assume that $p(+|C) = 0.9$ and $p(+|NC) = 0.08$. For $\epsilon = 0.15$ show that we now need at least 17 positive test results to conclude with 99% probability that a patient has cancer.

We can determine the probability of n positive tests for a cancer status κ by multiplying together the individual conditional probabilities,

$$p(+^n|\kappa) = \prod_{m=1}^{n} p(m+|(m-1)+,\kappa) \tag{3.97}$$

$$= \prod_{m=1}^{n} \left[1-(1-p(+|\kappa))e^{-(m-1)\epsilon}\right] \tag{3.98}$$

We can then determine the probability that an individual has cancer given n positive test results,

$$p(C|+^n) = \frac{p(+^n|C)\times p(C)}{p(+^n|C)\times p(C)+p(+^n|NC)\times p(NC)} \tag{3.99}$$

$$= \frac{p(C)\prod_{m=1}^{n}\left[1-(1-p(+|C))e^{-(m-1)\epsilon}\right]}{p(C)\prod_{m=1}^{n}\left[1-(1-p(+|C))e^{-(m-1)\epsilon}\right]+p(NC)\prod_{m=1}^{n}\left[1-(1-p(+|NC))e^{-(m-1)\epsilon}\right]} \tag{3.100}$$

If we graph this function we see that it takes many more tests to reach the required threshold (Figure 3.4), which is obtained after 17 tests.

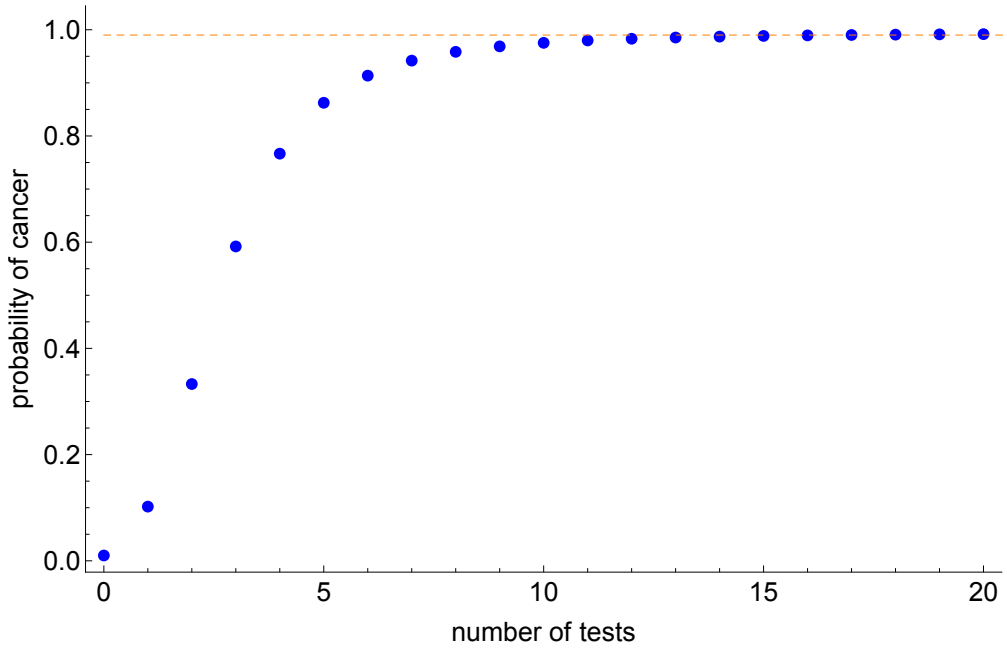

Figure 3.4: A plot of the probability of cancer as a function of the number of tests when we allow for persistence in the test results.

Bibliography