Bayesian Analysis of Intratumoural Oxygen Data

by

Herbert Hoi Chi Tang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Applied Mathematics

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

There is now ample evidence to support the notion that a lack of oxygen (hypoxia) within the tumour adversely affects the outcome of radiotherapy and whether a patient is able to remain disease free. Thus, there is increasing interest in accurately determining oxygen concentration levels within a tumour. Hypoxic regions arise naturally in cancerous tumours because of their abnormal vasculature and it is believed that oxygen is necessary in order for radiation to be effective in killing cancer cells. One method of measuring oxygen concentration within a tumour is the Eppendorf polarographic needle electrode; a method that is favored by many clinical researchers because it is the only device that is inserted directly into the tumour, and reports its findings in terms of oxygen partial pressure (PO2). Unfortunately, there are often anomalous readings in the Eppendorf measurements (negative and extremely high values) and there is little consensus as to how best to interpret the data. In this thesis, Bayesian methods are applied to estimate two measures commonly used to quantify oxygen content within a tumour in the current literature: the median PO2, and Hypoxic Proportion (HP5), the percentage of readings less than 5mmHg. The results will show that Bayesian methods of parameter estimation are able to reproduce the standard estimate for HP5 while providing an additional piece of information, the error bar, that quantifies how uncertain we believe our estimate to be. Furthermore, using the principle of Maximum Entropy, we will estimate the true median PO2 of the distribution instead of simply relying on the sample median, a value which may or may not be an accurate indication of the actual median PO2 inside the tumour. The advantage of the Bayesian method is that it takes advantage of probability theory and presents its results in the form of probability density functions. These probability density functions provide us with more information about the desired quantity than the single number that is produced in the current literature and allows us to make more accurate and informative statements about the measure of hypoxia that we are trying to estimate.

Acknowledgements

I would like to thank Rudy Gunawan, Dr. Giuseppe Tenti, and Dr. Sivabal Sivaloganathan who all helped me immensely on this thesis. I would also like to thank Dr. Michael Milosevic for providing the data used in this thesis.

Dedication

This is dedicated to my parents, without whom I would not be where I am right now.

Contents

List of Figures

Chapter 1

Introduction

Until the computer became an affordable and prevalent tool for the average researcher, Bayesian methods for analyzing data were often dismissed, ignored, and relegated to obscurity because they required large amounts of computational power, and because it was viewed (incorrectly) that Bayesian methods were vague and subjective. With computers becoming cheaper and more powerful by the day, and the advantages of the Bayesian method apparent, there is currently a renewed and escalating interest in using Bayesian methods to analyze data in many diverse research areas, including physics, chemistry, finance, computer science, and medicine. For example, Bayesian methods are currently being used by Berry [4] and his colleagues at the M.D. Anderson Cancer Center in Houston, Texas to design clinical trials; in fact, the Food and Drug Administration in the United States have discussed extensively in recent years how Bayesian methods might be adopted to improve the efficacy of clinical trials in general [5][6]; Bretthorst [3] (a chemist) uses it to analyze spectral data; Sivia [2] (a physicist) describes in detail applications in image processing in the latter chapters of his book; Jaynes [1] (also a physicist) builds his book on Bayesian methods upon the premise of teaching a robot how to learn (a field now known as machine learning); and in chapter 5 we will discuss how Bayesian methods are currently being used in email spam filters. In this thesis, the primary goal will be apply Bayesian methods to estimate two measures that clinical researchers use to quantify the amount of oxygen present inside a cancerous tumour. The need for accurate determination of oxygen content in tumors has become more important in recent years because there is increasing evidence that a lack of oxygen in tumours adversely affects the outcome of radiotherapy and increases the likelihood of the cancer migrating to other areas of the body.

Probability theory is key in data analysis, and many objected to the Bayesian interpretation of probability as a *degree of belief* or *measure of uncertainty* as they felt it was a definition that was imprecise and subject to personal bias. Instead, mathematicians developed what we will call orthodox or Frequentist statistics, the 'ordinary' statistics that is usually taught at the undergraduate level, where probability is defined as the frequency of occurrence of the event of interest over a large number of (infinitely many) experimental trials. This interpretation of probability is deemed to be both objective and measurable; probability is now an intrinsic physical property of the system and there is no room for personal bias. However, the frequency interpretation is also limited because not all problems involve events that occur randomly in repeated experimental trials. Sivia [2], for instance, tells of how Laplace was able to estimate the mass of Saturn to a high degree of accuracy (modern estimates differ by just 0.63%), using orbital data and celestial mechanics, by appealing to what we would call Bayesian methods. On its face value, Frequentists would not be able to tackle such a problem since mass is constant and not a random variable. Yet, we must be able to solve data analysis problems when the quantity of interest is not a random variable, otherwise, we would be helpless in many real life situations. So there needs to be some kind of resolution. One way of resolving the conundrum is to imagine an infinite number of parallel universes that were identical to our own except for the mass of Saturn and posit that our data consists of random samples from this ensemble of parallel universes. In such an alternate reality, mass would be a random variable and there is no contradiction in using Frequentist methods but clearly such a proposition is absurd. Although we are free to envision any scenario we wish to satisfy the frequency interpretation, the mass of Saturn does not change each time we try to estimate it, so Frequentist scholars usually do not resort to such intellectual games. Rather, they have invented numerous ad hoc methods that remove the logical inconsistency on a case by case, problem by problem basis.

For parameter estimation (such as in estimating the mass of Saturn) it is often assumed that the random variable is the error incurred when making measurements during repeated experimental trials. For most purposes, that is an acceptable compromise and makes sense logically, but in the case of Laplace, collecting orbital data is not the same as performing repeated experiments in the lab, the positions of the planets do not reset themselves so you can perform the same measurement again, and so there is still a disconnect between real life and orthodox statistics. The advantage of the Bayesian method is that we do not need to force ourselves into

giving every problem a frequency interpretation. Treating probability as a degree of belief agrees with our intuition, it allows us to avoid the ad hoc methods that are sensible for some situations but quite unintelligible for others, and we shall see that such an interpretation is not subjective except under perverse, mathematically unverifiable conditions.

To understand why Bayesians interpret probability as a degree of belief, we need to realize that by data analysis what we really mean is logical inference: the ability to make an informed conjecture about some quantity of interest based on observed data and relevant previous knowledge. Suppose you were outdoors, relaxing on the grass at a nearby park say, and nowhere near any radios, televisions, blackberries, iphones etc that could give you an up to date weather report. How do you know whether it will rain or not, and possibly ruin your day? The first thing that you would probably do would be to look up at the sky, and see whether there are any ominous dark clouds approaching; you might also consider the temperature, and whether the air felt damp or not. Further, you might even try to see if any flocks of birds are fleeing the area. Then, perhaps, you will remember the weather report that you saw before leaving the house, or maybe you will wonder whether it usually rains during this particular time of year or not. After taking into account all this information, which includes observed data (the clouds, the temperature, the birds) and previous knowledge (the earlier weather report, knowledge about seasonal trends), you would be then able to make an educated guess by performing some kind of logical reasoning in your mind (e.g. it will not rain because there is not a cloud in the sky and because the weatherman said it would be nice and sunny today). You might be wrong (dark clouds suddenly appear, you actually misheard the weather report for a different city) but it would be an opinion formed using the available information and we could not ask for more.

This act of making an estimate based on pertinent information is what we mean by logical inference, and it is exactly what happens in data analysis. When CNN collected poll data on how people felt about Obama before the election, they were trying to guess whether Obama would be eventually elected. When the World Health Organization collects data on the swine flu, they are trying to ascertain whether it will turn into a pandemic. When you measure something with a ruler, you are trying to estimate its length. The process of making an informed judgement based on available observations and prior knowledge is in fact quite general and prevalent in our everyday lives. The reality is that the nature of our existence

as humans precludes us from information other than what we can observe and what we have observed already. Most of us have intractable probability measures in our heads that allow us to take advantage of this information to make everyday decisions, and Bayesian methods attempt to quantify that process in order to solve data analysis problems.

Notice that in all the preceding examples the best we can do is offer a guess, a *probability* if you will, and we are unable to say with absolutely certainty one way or another (though we may be fairly convinced that we are correct). We will call such a guess our *degree of belief* in whatever we are estimating because humans are capable of believing certain propositions more than others, depending on the evidence present. Returning to our rain example, we would be more likely to believe that it would rain if we saw dark clouds in the sky than if we did not, so our degree of belief in the proposition that it will rain today would be higher in the former than in the latter. Intuitively, such reasoning is what we mean by probability, when we see that clouds in the sky, we might remark that it will *probably* rain today, or we might find it *more plausible* if one of our friends speculated as such, which is why Bayesians interpret mathematical probability as a degree of belief, and not as a frequency.

Frequentists object that interpreting probability as a degree of belief inherently makes it subjective because one person may believe one thing while the next person believes in another in the same situation. For instance, you might be convinced that it will rain, while your friend insists that the clouds will pass. The possible subjectivity of the Bayesian interpretation is certainly a valid concern, but if we examine the situation more closely we will see that it is unfounded. The key is in the information, by which we mean observable data and prior knowledge. Bayesians demand that two people having the same information come to the same conclusion. For example, in a data analysis setting, we require that objectivity mean that two different researchers with the exact same information come to the same conclusion. In particular, we will see in chapter 7 and Appendix E that Bayesians require information to be *testable*, which loosely means information that is (mathematically) verifiable, so as to eliminate situations where subjectivity might be an issue.

The inability to make definitive conclusions is what separates induction from deduction, and what makes induction more difficult. If we knew it was going to rain, then we would deduce that there would certainly be dark clouds in the sky.

However, if we observe that there are dark clouds in the sky, we would find it *more plausible* that it would rain, but it might not, and we would not know for sure unless we waited to find out. Orthodox statistics is deductive, it turns the problem of induction around and asks what if? For example, in hypothesis testing, we ask ourselves: what if the null hypothesis is true? What then is the probability of obtaining the observed data? Such a question is certainly a valid one, and we arrive at nice concrete conclusions, but those conclusions tells us very little because they ignore many other possible explanations. What happens when the null hypothesis is false? Bayesian methods, on the other hand, are inductive, they take advantage of probability theory - particularly Bayes' theorem - and try to make as few assumptions as possible about the underlying processes in producing a best guess at the quantity of interest. The conclusions will seem more vague at first glance, they will necessarily have to cover a wide range of possible answers, but they are in fact more truthful and more accurate, and the methods will be more generally applicable and more easily adaptable than ones in orthodox statistics.

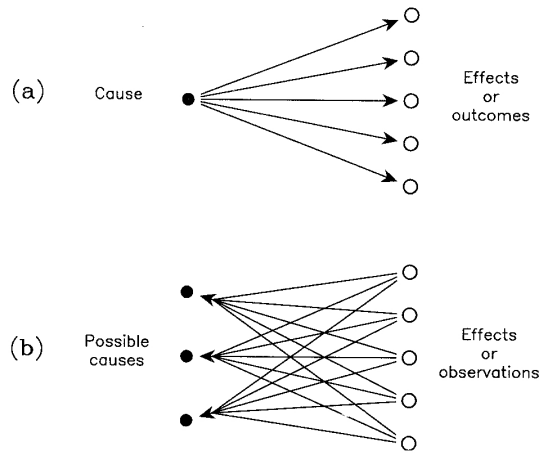


Figure 1.1: A schematic drawing showing the difference between deductive and inductive logic. The former is straightforward and concrete, given a definite set of premises, we can determine the set of possible outcomes without any ambiguity. However, if we are only given the outcomes, it is not so obvious what led to those observations. Figure from Sivia [2]

Moreover, Bayesian methods place a strong emphasis on incorporating prior information, whereas in orthodox statistics prior information is often an afterthought or footnote to be considered vaguely in explaining their results (hence the name ex-

planatory variables), and that will lead to better and more comprehensive analyses. A good example of this is sample size. In many areas of everyday life, baseball for example, one will often read comments such as "due to the small sample size the conclusions are not reliable" without much further qualification. We then might naturally ask: how unreliable is the conclusion? What constitutes a large enough sample size? In orthodox statistics, there is no clear answer to either of those questions; it is largely a matter of personal opinion. However, for Bayesian methods, we will often be able to quantify the accuracy of our conclusions succinctly with a measure called the error bar, a quantity that varies with sample size, so that there is no confusion about the reliability of our conclusions. If the sample size is small, the error bar will be large, but we will still be able to make definite statements about the quantity of interest.

Unfortunately, the inclusiveness of the Bayesian methodology is what makes it computationally expensive and devising algorithms to improve the efficiency of the numerical computation for Bayesian methods is an active research area. The deductive methods in orthodox statistics are designed to reduce to the same few standard distributions, so they can avoid computation all together and simply look up the values in tables of precalculated probabilities. This ease in computation is what made Frequentist methods preferable and convenient before the widespread use of computers.

It may seem in the preceding discussion that the Frequentist and Bayesian interpretations of probability, and hence their methods, are incompatible but that is not the case. Rather, the frequency interpretation is a special case of the Bayesian interpretation, and often the Bayesian methods will look remarkably similar to the Frequentist methods in situations where a frequency interpretation is valid and useful. Least squares and maximum likelihood, for example, are techniques that appear in both schemata but the interpretation of the results is different, and the Bayesian versions yield more information. In particular, we will see in chapter 6 how the Bayesian version of maximum likelihood leads to the aforementioned measure of accuracy known as the error bar; an extra tool that is not available in orthodox statistics.

As mentioned earlier, this thesis will focus on the application of Bayesian methods to estimate tumour oxygenation. We have been working with Fyles et al [11], a group of clinical researchers at Princess Margaret Hospital, and they have pro-

vided data from a probe that purportedly measures oxygen concentration inside a tumour. Specifically, we have oxygen partial pressure readings from various points inside the tumour for 105 cervical cancer patients. Tumour vasculature is highly irregular and that usually leads to a lack of oxygen, called hypoxia, in many areas of the tumour. Chapter 2 will examine how this irregular vasculature comes about and how hypoxia is believed to adversely affect radiotherapy.

The probe used to measure oxygen content in a tumour, is known to chemists as a polarographic oxygen sensor. Most clinical researchers around the world use such a probe made by the Eppendorf AG company [12][14][15][16][17][18]. The probe is a device that was originally used to measure concentrations of oxygen and oxygen reducible substances in chemical solutions and has since been adapted to be used in tumours, with mixed results. Chapter 3 investigates the theoretical workings behind the probe and potential issues that arise when using the probe in biological tissue.

In chapter 4 we will closely examine the actual data and consider the way the oxygen probe data is analyzed in the current literature. In particular, we will define a measure of oxygen content known as hypoxic proportion, the percentage of readings under a certain threshold value, and see if that measure is able to predict whether a patient will remain disease free. By disease free we mean the patient surviving and not suffering a relapse of the original cancer. Methods of analyzing survival, the Kaplan-Meier estimator and the log-rank test, will be presented in detail.

We have discussed at length already how Bayesians interpret probability as a degree of belief, and in chapter 5 we will see how that piece of insight is enough to derive the mathematical framework necessary for logical induction. The discussion will be grounded first by examining how Bayesian spam filters work to eliminate spam emails from our inboxes and it will become clear how probability theory is helpful in problems of data analysis and how it naturally arises from considerations in logical inference. Appendix E extends the discussion of Bayesian methods by presenting one of its more famous tenets: the idea of maximum entropy. We just mentioned that Bayesians make use of probability theory, but how do we assign these probabilities? Many scholars have argued that given some *testable* information, maximum entropy is in some sense the best and most objective way of assigning probabilities that reflect the given information and we will see a heuristic

justification of the maximum entropy principle by considering a hypothetical team of monkeys.

Chapter 6 is one of the main results of this thesis, the Bayesian estimate of hypoxic proportion, and chapter 7 is the other: the maximum entropy estimate of the median value of the distribution for the partial pressure of oxygen inside the tumour. Although hypoxic proportion is now the most commonly used measure of oxygen content, the median value is also almost always reported and compared to the the hypoxic proportion as a way of confirming the validity of hypoxic proportion as an indicator of tumour oxygenation, so it is of interest to estimate both quantities and to see if and how the Bayesian estimate differs. Moreover, the median value was the chief measure of oxygenation before researchers invented the notion of hypoxic proportion, and the sample median is sometimes chosen as the threshold value.

The final chapter, chapter 8, is a proposed preliminary alternative to the Kaplan-Meier estimator in estimating survival probability using an inhomogeneous Poisson process. Poisson processes are used in many practical situations where counting is involved, such as the number of customers arriving at a store within a certain time period, or the number of cracks over certain sections of a railway track, and, if we count the number of death or relapses over time, then survival can also be modeled under a Poisson process.

A comparison of my results to those of the existing literature will demonstrate the advantages of the Bayesian method over orthodox statistics, and provides reason why further work in this direction should be pursued. In particular my method of estimating hypoxic proportion not only reproduces the standard result in the literature but provides us with an estimate of how accurate that result may be. Moreover, the simplicity of the Bayesian method and its emphasis on prior information allows us to refine that estimate if more data or information become available. The maximum likelihood estimate of the median is more interesting, and yields values that are completely different from the median in the data. That the values are different is expected and welcome because the sample median is very likely not the median of the actual distribution of oxygen inside the tumour.

Chapter 2

Radiotherapy and The Role of Oxygen

Radiotherapy is a commonly administered treatment option when faced with many types of cancers. The American Society for Radiation Therapy estimates that around two thirds of all cancer patients will receive radiotherapy of some form during their treatment programs [43]; thus it is of interest to doctors and radiologists to investigate any factors that compromise or enhance the effectiveness of radiation therapy since even minor improvement in treatment strategies can benefit a large number of patients. It has been observed since the beginning of the twentieth century [9] that the oxygen concentration levels within a tumour have a significant impact on the effectiveness of radiation. In fact, Harrison et al [10] claim that oxygen may be "the most important determinant of response [to radiotherapy] among tumors of the same type". There is now an accepted theory as to why that is the case [9] and much corroboratory evidence to support the notion that the lack of oxygen in hypoxic tumours adversely affects the outcome of radiotherapy [10][11][12][14][15][16][17][18], but analyses tend to be inconclusive because methods used to measure the oxygen concentration only sample a small area of the tumour, different researchers do no agree on how best to interpret the data, and because conventional statistical methods used to study the data are inconclusive. Thus, there is much work to be done in the area. For now, we will focus on the mechanisms behind radiation therapy and how oxygen, or rather the lack of oxygen, in tumours affects radiotherapy.

2.1 Hypoxic Regions in Tumours

A basic question that needs to be addressed is why there are almost always regions of low oxygen, known as hypoxic regions, in any kind of solid tumour. Malignant tumours are marked by the rapid and uncontrolled proliferation of mutated cells. These mutated cells, like any other cell, need oxygen and nutrients to survive and so they will express pro-angiogenesis factors in an attempt to promote the growth of new blood vessels towards themselves to supply the necessary materials. However, because these cells are rapidly proliferating, they will cause an imbalance of pro- and anti-angiogenesis factors that leads to growth of abnormal, poorly formed, leaky blood vessels that do not do a good job of supplying well oxygenated blood to the entire tumour. Moreover, because tumour cells tend to be tightly packed together, blood vessels can become compressed and further impede the flow of blood to certain areas inside the tumour [41] (See figure 2.1). As a result, tumour vasculature is highly heterogeneous, and there are likely to be regions where cells do not receive adequate oxygen, causing them to become quiescent or inactive; and there may even be regions, usually near the centre of the tumour, where cells have died because of the lack of oxygen, known as the necrotic core [42] (See figures 2.2 and 2.3).

The kind of hypoxia described in the previous paragraph is known as "chronic hypoxia"; that is, it is persistent and long lasting. Another kind of hypoxia, called "acute hypoxia" is caused by a temporary closing of blood vessels in the tumour, and according to Hall [9] that happens because blood vessels in tumours apparently randomly open and close. Such transient regions of hypoxia are difficult to track, and it is not possible to make a distinction between the two from the data that we will be analyzing.

2.2 General Overview of Radiotherapy

Radiotherapy functions by using ionizing radiation to kill cells (preferably just cancer cells). It can be used in combination with other treatments, such as chemotherapy or surgery, or it may be the only prescribed course of treatment. How radiation therapy is administered varies from patient to patient. There are many different methods for focusing the radiation onto the tumour to minimize the side effects, but the mechanism behind radiotherapy remains the same.

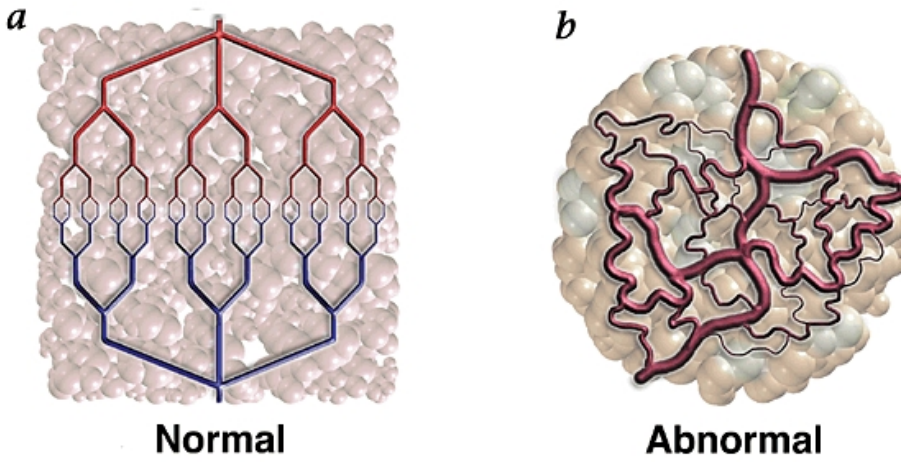


Figure 2.1: A comparison between normal vasculature (a) and tumour vasculature (b). Vasculature in normal tissue have regular networks that are made up of "mature" blood vessels, blood vessels that are in some sense ideally suitable for blood transport and well maintained by a balance of pro- and anti-angiogenic factors. Tumour vasculature is a chaotic network of blood vessels of varying diameters, varying lengths, and varying composition caused by an imbalance of the pro- and anti-angiogenic factors. This lack of regular structure causes some areas of the tumour to not be able to access an adequate supply of blood, and hence oxygen. Figure from [41]

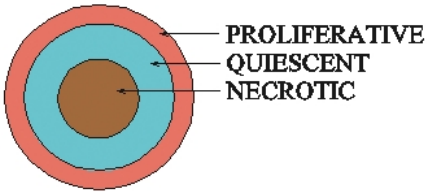


Figure 2.2: A simple diagram illustrating how cells are generally distributed within a tumour as a result of the abnormal vasculature. The appearance of both the necrotic cells (dead cells), and the quiescent cells (inactive cells) in tumours are due to the lack of oxygen reaching (usually) the centre of a solid tumour. Figure from [42].

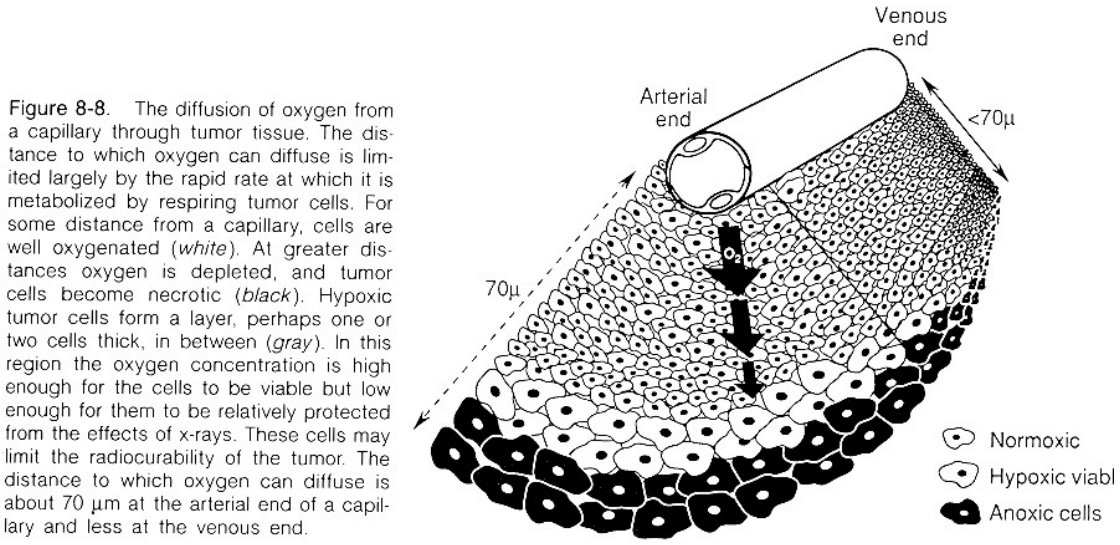


Figure 8-8. The diffusion of oxygen from a capillary through tumor tissue. The distance to which oxygen can diffuse is limited largely by the rapid rate at which it is metabolized by respiring tumor cells. For some distance from a capillary, cells are well oxygenated (*white*). At greater distances oxygen is depleted, and tumor cells become necrotic (*black*). Hypoxic tumor cells form a layer, perhaps one or two cells thick, in between (*gray*). In this region the oxygen concentration is high enough for the cells to be viable but low enough for them to be relatively protected from the effects of x-rays. These cells may limit the radiocurability of the tumor. The distance to which oxygen can diffuse is about 70 μm at the arterial end of a capillary and less at the venous end.

Figure 2.3: One of the main reasons why there are hypoxic regions inside a tumour with abnormal vasculartur is because the cells are too far away from the blood vessel. Figure from [9].

2.2.1 Ionizing Radiation

Ionizing radiation is radiation that has enough energy to cause one or more orbital electrons to be ejected by the atom or molecule that absorbs the radiation. For external beam radiotherapy, the most commonly used form of radiotherapy, the radiation is usually generated by using a linear accelerator to excite a beam of particles, such as photons, neutrons, or electrons, in the direction of the region that is to be irradiated.

2.2.2 Cell Killing

Ionizing radiation kills cells by damaging their DNA. If the damage is severe enough (more details can be found in [9]), the cells will no longer be able to divide and proliferate so they simply die the next time they try to do so. However, cells do have limited DNA repair mechanisms, so the radiation may not have any effect, or may just slow down the rate of proliferation, possibly inducing damage to subsequent daughter cells that make them more susceptible to radiation or more likely to die. It is believed that cancer cells are less likely to be able to repair the DNA damage caused by radiation than normal cells, one reason why there are usually minimal side effects when undergoing radiotherapy.

2.2.3 Direct and Indirect Action of Radiation

Radiation can damage DNA either directly or indirectly (see figure 2.4):

- **Direct action** means that the biological material (in our case the DNA) is absorbing the radiation directly, creating ionized or excited atoms in the material itself that eventually lead to biological change (e.g. mutation or death).
- **Indirect action** means that the radiation is interacting with the surrounding material and not the DNA itself. Such interactions can lead to the formation of free radicals that are highly reactive and have extremely short lifetimes (about 10^{-10}s). These highly reactive particles can then travel over and cause permanent damage to the DNA provided that they are produced within about 4nm of the DNA strands.

The indirect action of radiation tends to dominate for the types of radiation we normally use for radiotherapy, such as x-rays and γ-rays, and Hall [9] claims that for x-rays roughly two thirds of damage to DNA is caused by such indirect interactions.

2.2.4 The Oxygen Effect - Oxygen Fixation Hypothesis

As mentioned previously, cells are capable of limited DNA repair, so in order to kill the cell we need the damage to be permanent and not repairable. For indirect action, oxygen is said to "fix" the damage in place because it is believed that if, and only if, oxygen is present in the cell then the free radicals created by the indirect action of radiation will bond with the oxygen and create a product that is able to cause the desired permanent damage. Thus, a well oxygenated tumour is crucial for radiation therapy to be effective, and there is experimental evidence that hypoxic cells are two to three times more resistant to radiotherapy [9].

2.2.5 Fractionation

Recall that in a tumour there are likely to be pockets of hypoxia where cells are not well oxygenated because tumour vasculature is highly chaotic and irregular. Thus, there is a somewhat paradoxical balance between cancer cells which are more susceptible to radiation in general and hypoxic cancer cells that are more resilient and able to counteract its effects. The presence of these hypoxic cells means that

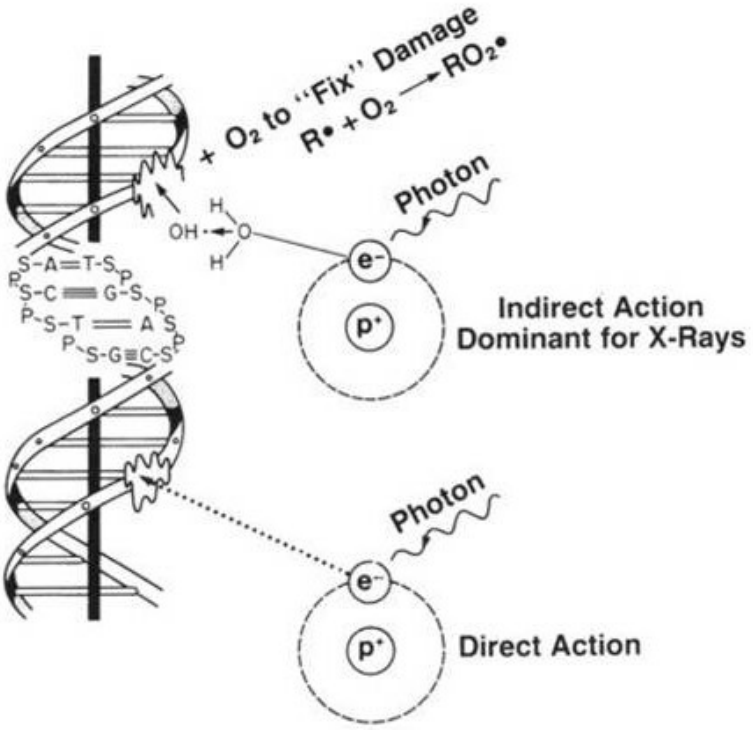


Figure 2.4: The direct and indirect action of radiation on DNA. For indirect action, the most common free radical is the hydroxyl radical, OH·, because 80% of a cell is water and that is the molecule that is produced when water is ionized. For any free radical, R·, once it has diffused to the DNA molecule and caused damage, a subsequent reaction with oxygen will then "fix" the damage in place, i.e. make it permanent and not repairable, leading eventually to cell death. Diagram from [9].

a single large dose of radiation is not effective; it is more likely that healthy cells will be damaged than the hypoxic cells, and hence likely to cause more harm to the patient than desired. Rather, it is better to apply several small doses, called fractions, killing off the well oxygenated cancer cells each time, and allowing the hypoxic cells access to oxygen that that was not available previously. So that, after some time, they will become well oxygenated, in a process called "reoxygenation", and become more susceptible to radiation for subsequent treatments. The more hypoxic a tumour is, the less effective radiation will be, and that is why doctors are interested in assessing oxygen concentration in tumours.

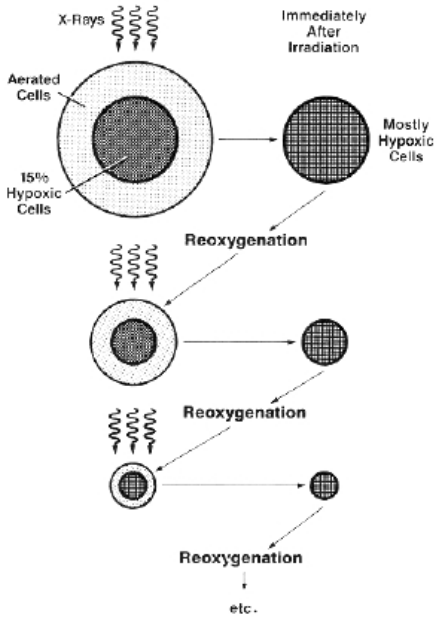


Figure 2.5: Applying one large dose of radiation is extremely harmful to healthy cells and unlikely to eliminate the tumour because the hypoxic cells in the tumour will resist the radiation. So it is better to administer several small doses, allowing the previously hypoxic cells to "reoxygenate" and become more susceptible to radiation in subsequent treatments. Notice that we don't care about the necrotic core because those cells are dead already, so they make no difference. Diagram from [9].

2.3 Metastasis

Hypoxia has also been implicated in playing a role in metastasis, the spread of cancer from the initial site to other areas of the body [10], [11]. In particular, hypoxic tumours seem to preferentially select cells with mutated p53 genes, a gene that normally suppresses tumour growth, so they become more aggressive and more likely to metastasize [10]. For example, Fyles et al [11] found that a hypoxic tumour for their cervical cancer patients was a good indication that the cancer had metastasized to the pelvic lymph nodes. In short, accurately determining the amount of oxygen in tumours is not only important for radiotherapy, but could have other implications as well.

2.4 Methods for measuring oxygen concentration in tumours

There are two current ways to measure the amount of oxygen in a tumour: biomarkers (such as pimonidazole) and polarographic probes. Both only sample parts of the tumour and give, either directly or indirectly, a measure of how much we believe a tumour to be hypoxic. However, there seems to be little agreement between the two [19], which is troubling. For the rest of this thesis, I will focus on how the polarographic probe works and how we might analyze and interpret the measurements taken by the probe.

Chapter 3

An Overview of The Polarographic Oxygen Electrode

Physicians and medical researchers now routinely measure tumour oxygenation using a device known as a polarographic oxygen sensor. It is an instrument that was originally developed to measure concentrations of oxidizable and reducible substances in solutions, and now has been adapted and miniaturized to measure the partial pressure of oxygen in tissue. Most clinical researchers use a model made by Eppendorf AG in Germany, so we will often refer to the probe simply as Eppendorf. We will also sometimes refer to oxygen partial pressure as oxygen tension as that is the common term in medical literature. Whether a polarographic probe actually measures partial pressures in tissue is debatable since the theoretical principles upon which the probe is based rely on assumptions that probably do not hold in tissue, but it is generally accepted that polarographic probes are able to measure concentrations of dissolved ions that are oxygen reducible or oxidizable in solutions [24]. When we examine the actual data in chapter 4 we will see that the probe gives us many anomalous readings (such as negative or extremely high partial pressures) that lends credence to the notion that the probe may not be reliable. However, it is currently the only instrument that actually interacts with the tumour microenvironment directly, so it is still believed to be the best available method of assessing tumour oxygen concentration despite some concerns.

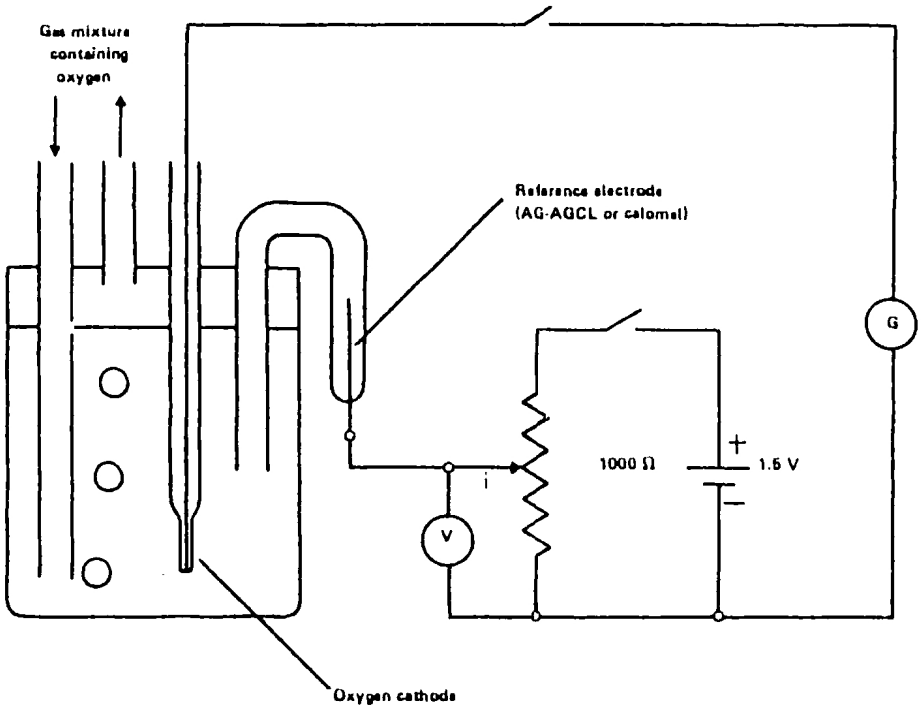


FIGURE 1. Diagram showing an experimental apparatus that will measure cathodic current due to dissolved oxygen. The reference electrode can be either a calomel or silver-silver chloride half-cell. The 1,000-ohm potentiometer permits adjustment of the potential between the cathode and reference electrode. (After Davies, 1962.)

Figure 3.1: Simple circuit diagram of the idea behind a polarographic oxygen sensor. Figure from [24].

3.1 Chemical Basis of the Probe

Briefly, a polarographic oxygen sensor consists of two electrodes inserted into the solution of interest with a potential difference applied across the electrodes. Oxygen reduction (redox) reactions will then occur at the cathode, and both Fatt [24] and Hahn [29] claim that the most likely one is the following (especially if the probe is covered by a membrane):

$$O_2 + 2H_2O + 4e^- = 4OH^-. \tag{3.1}$$

Gold or platinum is usually chosen for the cathode so as to avoid other reactions that may occur that could affect the current and make it more likely that only the above sequence occurs. To ensure that the above reaction does not depend on the solution, a known reaction is chosen for the anode so that it acts as a "reference" that fixes the voltage across the cathode [24]. The chosen reaction is usually silver-silver chloride:

$$Ag + Cl^- = AgCl + e^-. \tag{3.2}$$

Together, the above two reactions will generate a current that depends on the amount of oxygen present in the solution. However, this current is also driven by the potential difference between the two electrodes, so it is favourable to find a region where the current is independent of the applied voltage, but is still enough to drive the reduction reaction. Fortunately, such a region exists, and in this region the electrode is said to be *polarized* because the current is largely unaffected by the external voltage applied onto it (see figure 3.2).

3.2 Mathematical Modeling of the Probe

3.2.1 Diffusion

For a general polarographic probe, where we allow the probe to measure the concentrations of any reducible or oxidizable ions, there are two main ways for those ions to reach the probe [26]:

1. Diffusion: If we assume that oxygen is immediately and completely consumed at the surface of the probe, then there will be a concentration gradient near the probe, driving more ions to enter the probe via diffusion.

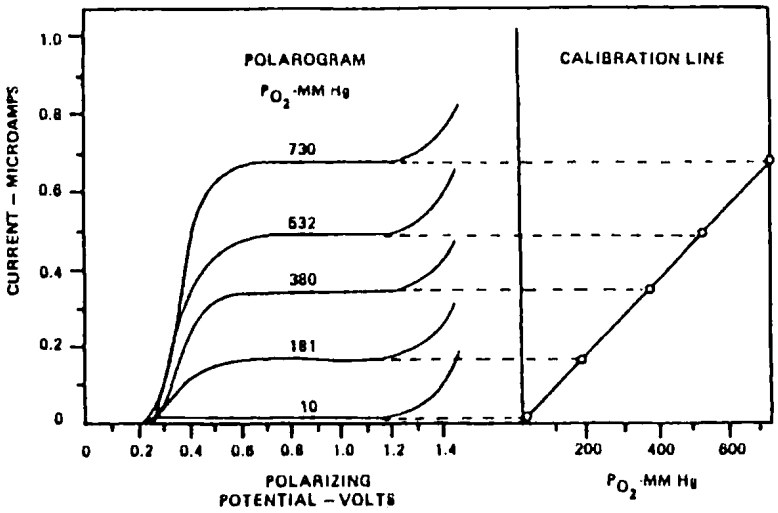


FIGURE 3. Cathodic current for a platinum electrode as a function of applied polarizing voltage in solutions of different oxygen tension. These curves represent a somewhat idealized situation because the plateau is often not as flat or well defined as shown here. (After Kimmich and Kreuzer, 1969.)

Figure 3.2: Notice between about -0.6V to -1.0V the current is approximately flat, i.e. independent of the applied voltage. That flat plateau is known as the polarized region. Figure from [24].

2. Electromotive force: Since we are applying a potential difference across our electrodes in order to drive the redox reaction, charged ions will also be affected by the electric field, and drift towards one of the electrodes. This is known as electrical migration.

In the case of charge neutral oxygen molecules, the second effect is negligible. Thus, we will assume that oxygen only reaches the probe via diffusion, and focus on how this relates to the current generated by the redox reaction. Some researchers, such as Mancy [27], suggest that we should also correct for other factors, such as salt concentrations that may be charged and affect the flow of oxygen molecules into the probe. Further discussion on this matter follows in section 3.2.5; for now we want to start with a simple model of the probe.

There are two main things to consider in our diffusion model:

1. How is current related to oxygen concentration/oxygen tension?
2. How quickly will the current reach a steady state?

The need to answer the first question is obvious. The second question is a matter of reproducibility. If the current generated is time dependent, then we would have to either take measurements at precisely the same instant each time we perform an experiment, or have to keep track of when the measurement is taken in order to be able to compare measurements later. Neither proposition is attractive as the first is difficult to achieve, and the second would probably involve complicated equations with many experimentally determined constants. To overcome this difficulty, recall that solutions to the diffusion equation eventually reach a time independent steady state, so it is in our interest to make the system reach this steady state as quickly as possible.

To answer the first question, we consider the usual diffusion equation,

$$\frac{\partial C}{\partial t} = D\nabla^2 C, \tag{3.3}$$

where C is the oxygen concentration and D is the diffusion constant. Then, we will apply Henry's Law,

$$C = kP, \tag{3.4}$$

to derive an equation that involves the partial pressure of oxygen instead of concentration. Substituting (3.4) into (3.3) yields:

$$\frac{\partial P}{\partial t} = D\nabla^2 P. \tag{3.5}$$

Many researchers, including Fatt [24], Hudson [30], and Albanese [28], simply assume that this law holds and use it without qualification. In reality, the above version of Henry's Law (3.4) holds only at thermal equilibrium, at a certain fixed temperature, and for low pressures (P<200kPa) and low concentrations (mole fractions of <0.01). Whether these conditions are satisfied in the context of biological tissues is debatable; for this section we will assume that the law holds so that we can work in partial pressures instead of concentrations. More discussion on why we want to work with partial pressures follows in section 3.3.3; in short, it is the convention to do so. Further details of modern corrections to Henry's Law are given in Appendix A. The diffusion equation will be solved subject to certain initial and boundary conditions depending on the situation we are trying to model, and these will become clear in the succeeding sections.

3.2.2 Equation for the Current

Once we solve for the partial pressure, we will need an equation that relates the partial pressure to the current. Fatt [24] proposes the following:

$$I = nFAf, \tag{3.6}$$

where

n = amount of electricity required for the redox reaction (in number of faradays)

F = the Faraday constant[1]

A = electrode area

f = the flux of oxygen into the electrode

[1] The Faraday constant is an old unit of charge that is frequently used in electrochemistry. It is defined as the charge per mole of electrons:

$$F = N_A e \tag{3.7}$$

where N_A is Avogadro's number, and e is the charge of an electron. F has a currently accepted value of 96 485.3399 C/mol.

Flux is given by Fick's law:

$$j = -D\frac{\partial C}{\partial x} = -Dk\frac{\partial P}{\partial x}, \tag{3.8}$$

where C is the concentration of oxygen, D is the diffusion constant, and k is the Henry's Law constant. It is not clear in any of the texts where Fick's law is mentioned [24][26][28] how j relates to f, but it appears that f in the current equation is set equal to either $-j$ or $|j|$ so that the flux, and hence current, is a positive quantity.

Although the equation makes sense intuitively, it is worth checking the dimensions of the quantities in the equation to ensure dimensional consistency:

$[I]$= Q/T

$[n]$= 1

$[F]$= Q/mol

$[A]$= L^2

$[f]$= mol/L^2/T

Thus, the equation is dimensionally consistent as expected.

3.2.3 Bare Spherical Electrode

For the partial pressure, we will simplify the problem by assuming spherical symmetry. Although most probes are not spheres, Davies [25] claims that the sphere is a good approximation because the "the diffusion field rapidly approaches spherical symmetry with distance from the tip". To begin, we will consider the situation where the probe is "bare" which simply means we will insert the probe directly into the tissue, as opposed to covering it first with a membrane to help protect it from unwanted chemical reactions (covered in the next section).

In spherical coordinates, assuming spherical symmetry, the diffusion equation becomes:

$$\frac{\partial P}{\partial t} = \frac{D}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial P}{\partial r}\right), \qquad r \geq a. \tag{3.9}$$

The equation only holds for $r \geq a$ because we assume that the probe has radius a. We will solve the above equation subject to the following conditions:

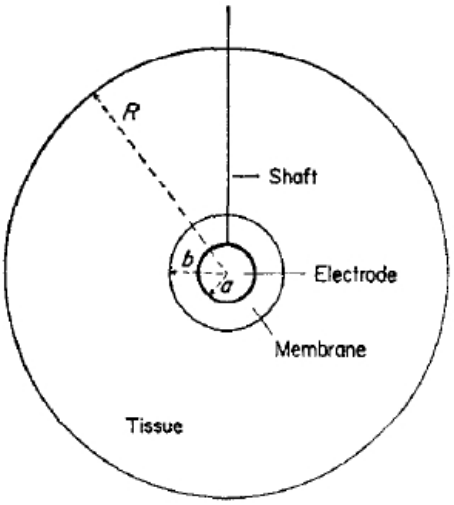


FIG. 1. Diagram of electrode-tissue model.

Figure 3.3: A schematic diagram of a spherical probe in tissue. We will assume for both the bare and membrane covered electrodes that the shaft is negligibly thin. This diagram is for the case of the membrane covered electrode; for the bare electrode, there is no membrane layer and we make no assumptions about the size of the tissue. Figure from [28]

- Initial condition: $P(r,0) = P_0$, for $r \geq a$.
- Boundary condition: $P(a,t) = 0$.

Clearly, the distribution of oxygen within normal biological tissue is not uniform, so the initial condition represents a sort of averaging of the actual conditions in tissue that we approximate to be constant so that the problem is easier to solve. The boundary condition at the surface of the probe reflects our idealization that oxygen is immediately and completely consumed as soon as it enters the cathode.

To solve for the partial pressure P, using techniques from Carslaw and Jaeger [33] appropriate for our problem, we make the change of variable $u(r,t) = rP(r,t)$ to transform (3.9) into:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial r^2}, \qquad r \geq a, \tag{3.10}$$

subject to:

$$u(r,0) = rP_0, \qquad r \geq a, \tag{3.11}$$

$$u(a,t) = 0. \tag{3.12}$$

Now shift the axis: Let $x = r - a$, and $U(x,t) = u(r,t)$, then we have:

$$\frac{\partial U}{\partial t} = D\frac{\partial^2 U}{\partial x^2}, \qquad x \geq 0, \tag{3.13}$$

subject to:

$$U(x,0) = P_0(x+a), \qquad x \geq 0, \tag{3.14}$$

$$U(x,t) = 0. \tag{3.15}$$

Equation 3.13 is simply the semi-infinite rod problem where the end at $x = 0$ is kept in a cold bath, with initial condition $U(x,0) = f(x) = P_0(x+a)$. So to solve we make an odd extension:

$$f_{odd}(x) = \begin{cases} f(x) = P_0(x+a), & x \geq 0, \\ -f(-x) = P_0(x-a), & x < 0. \end{cases} \tag{3.16}$$

Then, using the Fourier transform method and applying the diffusion kernel we get:

$$U(x,t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi Dt}} e^{\frac{-(x-s)^2}{4Dt}} f_{odd}(s)ds, \quad x \geq 0. \tag{3.17}$$

We can simplify (3.17) into a more manageable expression as follows:

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi Dt}} e^{\frac{-(x-s)^2}{4Dt}} f_{odd}(s)ds &= \int_{0}^{\infty} \frac{1}{\sqrt{4\pi Dt}} e^{\frac{-(x-s)^2}{4Dt}} P_0(s+a)ds, \\
&\quad + \int_{-\infty}^{0} \frac{1}{\sqrt{4\pi Dt}} e^{\frac{-(x-s)^2}{4Dt}} P_O(s-a)ds \\
&= P_0 \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi Dt}} s e^{\frac{-(x-s)^2}{4Dt}} ds \\
&\quad + \frac{P_0 a}{\sqrt{4\pi Dt}} \int_{0}^{\infty} \left[e^{\frac{-(x-s)^2}{4Dt}} - e^{\frac{-(x+s)^2}{4Dt}} \right] ds.
\end{aligned}$$

Here we made a change of variable $s \to -s$ and reversed limits to get the second term in the second integral. The first integral is simply the expected value of a Gaussian:

$$P_0 \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi Dt}} s e^{\frac{-(x-s)^2}{4Dt}} ds = P_0 \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi Dt}} s e^{\frac{-(s-x)^2}{4Dt}} ds = P_0 x. \tag{3.18}$$

The second integral is trickier and requires a change of variable $z = (s-x)/\sqrt{4DT}$ in the first part and $z = (s+x)/\sqrt{4DT}$ in the second part:

$$\begin{aligned}
\frac{1}{\sqrt{4\pi Dt}} \int_{0}^{\infty} \left[e^{\frac{-(x-s)^2}{4Dt}} - e^{\frac{-(x+s)^2}{4Dt}} \right] ds &= \frac{1}{\sqrt{\pi}} \left[\int_{\frac{-x}{\sqrt{4Dt}}}^{\infty} e^{-z^2} dz - \int_{\frac{x}{\sqrt{4Dt}}}^{\infty} e^{-z^2} dz \right], \\
&= \frac{1}{\sqrt{\pi}} \int_{-\frac{x}{\sqrt{4Dt}}}^{\frac{x}{\sqrt{4Dt}}} e^{-z^2} dz, \\
&= \frac{2}{\sqrt{\pi}} \int_{0}^{\frac{x}{\sqrt{4Dt}}} e^{-z^2} dz, \\
&= erf\left(\frac{x}{\sqrt{4Dt}} \right).
\end{aligned}$$

Putting them together, we get:

$$U(x,t) = P_0 x + P_0 a \cdot erf\left(\frac{x}{\sqrt{4Dt}} \right), \quad x \geq 0. \tag{3.19}$$

Thus the solution to the original problem, (3.9), is:

$$P(r,t) = P_0\left(1 - \frac{a}{r}\right) + \frac{P_0 a}{r} erf\left(\frac{r-a}{\sqrt{4Dt}}\right), \quad r \geq a, \tag{3.20}$$

where

- a is the radius of the electrode,
- r is the distance from the center of the electrode,
- erf is the error function.

Now that we have an expression for $P(r,t)$ we can solve for the current by using (3.6), where the flux is given by $f = Dk(\partial P/\partial r)|_{r=a}$. Using (3.20) we get

$$\left.\frac{\partial P}{\partial r}\right|_{r=a} = P_0\left(\frac{1}{a} + \frac{1}{\sqrt{\pi Dt}}\right). \tag{3.21}$$

Since we assumed the electrode was spherical, we have $A = 4\pi a^2$, so the current is:

$$I = 4n\pi FDkP_0 a\left(1 + \frac{a}{\sqrt{\pi Dt}}\right). \tag{3.22}$$

Thus in the steady state we have

$$I = 4nFDkP_0\pi a. \tag{3.23}$$

This linear relationship between the partial pressure and current in the steady state is the key to all of polarography. It is only because of the linearity assumption that researchers can calibrate oxygen electrodes simply by using two different solutions of known concentration and not worry about having to find approximate values for all the constants (more on this later).

Notice that the only variable we have control over is the radius of the electrode a. According to Fatt [24], using a typical value of $D = 2\text{x}10^{-5} cm^2/\text{s}$, an electrode with a 0.5mm radius will reach within 10% of its steady state in about 4000s, whereas an electrode with radius 5μm will reach within 10% of its steady state in 0.4s. So a small probe is necessary for a reasonable response time, and luckily that is exactly

what we want in a needle electrode small enough to be inserted into human tissue. Several researchers also claim that a small probe reduces "stirring effects" which are anomalous currents caused by the solution moving past the probe [24], [26], [29], [31].

3.2.4 Membrane Covered Spherical Electrode

In biological tissue there are often a number of substances that can have an adverse effect on the probe, so in practise it is covered by an oxygen permeable membrane to try and ensure that only oxygen passes through to the cathode and to protect the electrode from other unwanted reactions. For the Eppendorf probe, the membrane is made of teflon [12].

To begin, we will assume that the probe, with radius a, has a layer of membrane surrounding it of radius b, and that the tissue surrounding the probe is homogeneous, with radius R. Since oxygen dissolves differently in each material, we will have to take solubility into consideration, and we will also attempt to model the fact that cells inside the tissue consume oxygen by adding a constant consumption term. Solving the diffusion equation for multiple layers is difficult, so we will only consider the steady state:

$$\alpha_M D_M \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial P}{\partial r}\right) = 0, \qquad a \le r \le b, \tag{3.24}$$

$$\alpha_T D_T \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial P}{\partial r}\right) - Q_T = 0, \qquad b \le r \le R, \tag{3.25}$$

where α_M, α_T are the solubility coefficients for oxygen in the membrane and tissue respectively; D_M, D_T are the diffusion coefficients of the membrane and tissue, and Q_T is the oxygen consumption rate of the tissue.

To solve equations 3.24 and 3.25 we follow the approach of Albanese [28] with slightly different notation. We begin by analyzing the situation before the potential is applied; this means that the electrode is not consuming any oxygen. The boundary conditions in this case are:

- $\frac{\partial P}{\partial r}\big|_{r=a} = 0$,
- $\alpha_M D_M \ \frac{\partial P}{\partial r}\big|_{r\to b-} = \alpha_T D_T \ \frac{\partial P}{\partial r}\big|_{r\to b+}$,

- $P(r)|_{r\to b-} = P(r)|_{r\to b+}$,
- $P(R) = P_R$, where P_R is known.

The first condition is a no flux boundary condition that tells us no oxygen is passing through into the electrode. The middle two constraints give matching conditions to ensure continuity of the solution, and the third is simply a fixed value of pressure at the other boundary that is assumed to be known.

Solving (3.24) and (3.25) is fairly straightforward and gives:

$$P(r) = \frac{A}{r} + B, \quad a \le r \le b, \tag{3.26}$$

$$P(r) = \frac{Q_T}{6\alpha_T D_T} r^2 + \frac{C}{r} + D, \quad b \le r \le R. \tag{3.27}$$

Applying the first boundary condition determines the first constant A:

$$\left.\frac{\partial P}{\partial r}\right|_{r=a} = -\frac{A}{a^2} = 0 \quad \Rightarrow \quad A = 0. \tag{3.28}$$

Since $A = 0$, $P(r) = B$ for the inner region, so the second condition now determines C:

$$\alpha_M D_M \left.\frac{\partial P}{\partial r}\right|_{r\to b-} = \quad 0 \quad = \alpha_T D_T \left.\frac{\partial P}{\partial r}\right|_{r\to b+}, \tag{3.29}$$

$$\Rightarrow \frac{Q_T}{3\alpha_T D_T} b - \frac{C}{b^2} = 0, \quad \Rightarrow \quad C = \frac{Q_T}{3\alpha_T D_T} b^3. \tag{3.30}$$

Plugging this value of C into the fourth condition gives us D:

$$D = P_R - \frac{Q_T}{6\alpha_T D_T} R^2 - \frac{Q_T}{3\alpha_T D_T} \frac{b^3}{R}. \tag{3.31}$$

Putting these all together into the matching condition at $r = b$ allows us to calculate B, which we will relabel as P_b, because this constant will show up again in the next derivation:

$$P_b := B = -\frac{Q_T}{6\alpha_T D_T} \left(R^2 - b^2\right) + \frac{Q_T}{3\alpha_T D_T} b^3 \left(\frac{1}{b} - \frac{1}{R}\right) + P_R. \tag{3.32}$$

Physically, this is the steady state partial pressure at the surface of the membrane, $r = b$, before a potential is applied on the electrode.

Once the probe is switched on, the first boundary condition changes as we assume all the oxygen is immediately consumed at the surface of the electrode:

- $P(a) = 0$,
- $\alpha_M D_M \left.\frac{\partial P}{\partial r}\right|_{r\to b-} = \alpha_T D_T \left.\frac{\partial P}{\partial r}\right|_{r\to b+}$,
- $\left.P(r)\right|_{r\to b-} = \left.P(r)\right|_{r\to b+}$,
- $P(R) = P_R$, where P_R is known.

Since our end goal is to obtain an expression for the current, which is proportional to the flux at the electrode surface, $\alpha_M D_M K \left.\frac{\partial P}{\partial r}\right|_{r=a} = -\alpha_M D_M K \frac{A}{a^2}$, we need only determine A explicitly.

Substituting equations (3.26) and (3.27) into the four boundary conditions above yields

- $\frac{A}{a} + B = 0$,
- $-\alpha_M D_M \frac{A}{b^2} = \alpha_T D_T \left(\frac{Q_T}{3\alpha_T D_T} b - \frac{c}{b^2}\right)$,
- $\frac{A}{b} + B = \frac{Q_T}{6\alpha_T D_T} b^2 + \frac{C}{b} + D$,
- $\frac{Q_T}{6\alpha_T D_T} R^2 + \frac{C}{R} + D = P_R$.

After some algebra we obtain

$$A = \frac{P_b}{\left(\frac{1}{b} - \frac{1}{a}\right) - \frac{\alpha_M D_M}{\alpha_T D_T}\left(\frac{1}{b} - \frac{1}{R}\right)}, \tag{3.33}$$

where P_b is as defined in (3.32). Thus, using the fact that the flux is $\alpha_M D_M K \left.\frac{\partial P}{\partial r}\right|_{r=a} = -\alpha_M D_M K \frac{A}{a^2}$, and substituting the expression into (3.6), the current is

$$\boxed{I = \frac{nFA^*\alpha_M D_T P_b}{a\left[\frac{D_T(b-a)}{D_M b} + \frac{\alpha_M}{\alpha_T}\left(\frac{a}{b} - \frac{a}{R}\right)\right]}}, \tag{3.34}$$

where A^* in this equation is the electrode area ($A^* = 4\pi a^2$).

Although this equation for the current preserves the linear relationship that we desire, it is a complicated expression that involves both the tissue and membrane properties. Albanese [28] notes that if

$$\frac{D_T(b-a)}{D_M b} \gg \frac{\alpha_M}{\alpha_T}\left(\frac{a}{b} - \frac{a}{R}\right), \tag{3.35}$$

then the current simplifies to

$$I = \frac{nFA\alpha_M D_M b P_b}{a(b-a)}, \tag{3.36}$$

which is independent of the tissue properties. Unfortunately this does not hold in general, so calibration is difficult because the constants depend on tissue properties, meaning a simple calibration procedure involving two solutions of known oxygen concentration in the lab would not give the correct slope or intercept in the linear relationship for a particular biological tissue.

Real tissue is not homogeneous, so Albanese goes on to consider the same spherical probe with an extra layer to demonstrate how difficult it is to remove the dependence on tissue properties for even just a slightly more complex model (Albanese suggests that this new layer might physically represent a clot around the electrode or some collagenous fibers). Equations (3.24) and (3.25) are now replaced by

$$\alpha_M D_M \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial P}{\partial r}\right) = 0, \qquad a \le r \le b, \tag{3.37}$$

$$\alpha_i D_i \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial P}{\partial r}\right) = 0, \qquad b \le r \le c, \tag{3.38}$$

$$\alpha_o D_o \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial P}{\partial r}\right) - Q = 0, \qquad c \le r \le R, \tag{3.39}$$

where the subscript i denotes the inner layer, $b \le r \le c$, and the subscript o denotes the outer layer, $c \le r \le R$. The current in this case is

$$I = \frac{nFA\alpha_M D_0 P_b}{a\left[\frac{D_0(b-a)}{D_M b} + \frac{\alpha_M D_o}{\alpha_i D_i}\left(\frac{a}{b}-\frac{a}{c}\right) + \frac{\alpha_M}{\alpha_0}\left(\frac{a}{c}-\frac{a}{R}\right)\right]}. \tag{3.40}$$

We could make a similar argument as before to simplify the denominator, but the expression is fairly complicated and difficult to satisfy physically. In general, the idea is that whilst we do have a linear relationship between current and pressure, the proportionality "constant" depends explicitly on tissue properties, so a calibration procedure to determine the constant experimentally will not work unless it is done in the tissue, which is either extremely difficult or impossible. The best we can say in this case is that we can detect relative changes, that is, increases or decreases in pressure with regard to other measured values, and that we do not determine the actual value of the pressure.

3.2.5 What is diffusing?

Short and Shell [32] claim that there is confusion in the literature over how to interpret the quantity C that appears in the diffusion equation:

$$\frac{\partial C}{\partial t} = D\nabla^2 C. \tag{3.41}$$

In most instances, such as Fatt [24], Albanese [28], and Kolthoff and Lingane [26], the quantity C is simply taken as the concentration, and no further justification is required because there is a clear concentration gradient caused by the probe that would lead to oxygen diffusing into the electrode. However, Mancy [27] claims that he only treats C as concentration in order to simplify the discussion, and that C should really be the chemical potential. The reason being that in solutions with high salt concentrations, the current generated by the probe is higher than it should be, indicating a higher oxygen concentration than is actually physically present. He says that we should correct for the activity of oxygen, because it is the activity that drives the redox reaction, and this is higher than the oxygen concentration in high salt concentrations; but Short and Snell [32] disagree, saying that it only makes sense for C to mean concentration because only concentration is a physical quantity that can diffuse.

Also, when we use Henry's law to convert from concentration to partial pressure, we get a diffusion equation for the partial pressure P:

$$\frac{\partial P}{\partial t} = D\nabla^2 P. \tag{3.42}$$

Physically, pressure is not a quantity that can diffuse, so the equation is misleading in the sense that pressure is not what is diffusing, it is the concentration that is diffusing, but since we assume that Henry's Law, $C = kP$, holds, the pressure must be varying at the same rate as the concentration, so mathematically the expression we get for pressure satisfies the same diffusion equation as for the concentration.

3.3 Practical Aspects of Polarography

3.3.1 Original uses in Chemistry

Polarography was first invented by Heyrovsky at Charles University in Prague in the 1920s [31], [26]. He used what is called a dripping mercury electrode in which

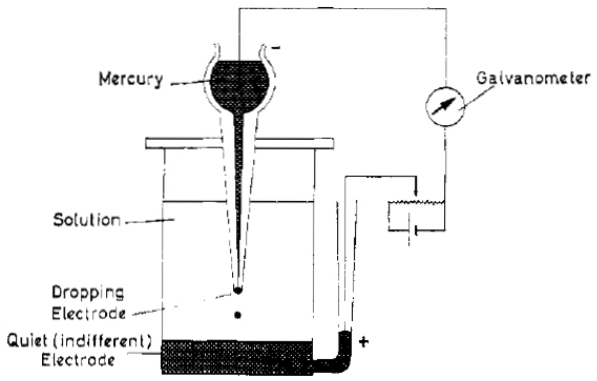


Fig. 1. Diagram of the dropping mercury electrode.

Figure 3.4: A diagram of one of the first kinds of polarographic probe. Figure from [31]

one electrode consisted of a glass tube that allowed tiny drops of mercury to fall onto the solution of interest connected to another electrode. Heyrovsky found that he got unique current-voltage curves for each electro-oxidizable or electro-reducible ion in the solution, and that it depended on the concentration of each substance. With this discovery, Heyrovsky and future chemists were able to determine the concentrations of different substances within a solution of unknown composition by examining the current-voltage curve generated by that substance and identifying the different curves for each substance.

Further developments would follow, starting with Heyrovsky himself and Shikata in 1925 when they designed an improved version of the dripping mercury electrode, which they called the polarograph, that recorded the results automatically. Later, the solid electrodes shown in figure 3.5 would follow, allowing for biological applications as it removed the use of mercury, which is poisonous to humans and harmful to biological tissue. According to Hahn [29], Clark type polarographic electrodes (see figure 3.5) are now used in food, alcohol, aircraft and space industries, and in soil chemistry, waste water, and sewage management.

3.3.2 Medical Applications

The polarographic probe is considered the "gold standard" for measuring tumour hypoxia by many medical researchers because it directly sample the tissue in vivo [11], [12]. However, Hahn [29], a researcher who uses similar probes for determining

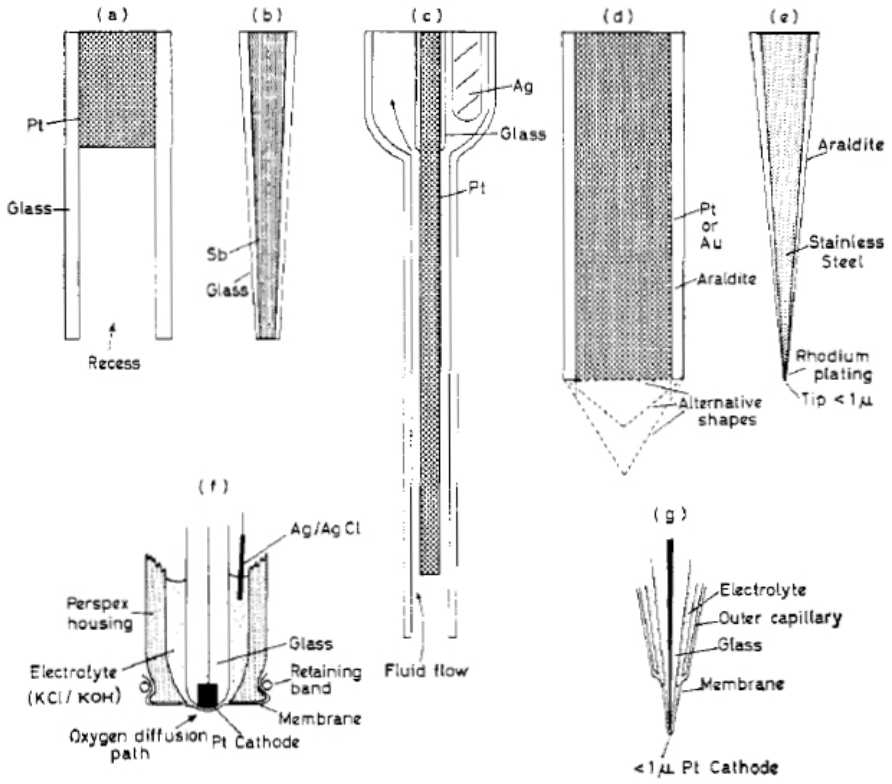


Fig. 2. Some oxygen cathode designs (not to scale).

(*a*) The recessed electrode (Davies and Brink 1942).
(*b*) The antimony micro-electrode (Davies and Rémond 1947).
(*c*) The flow electrode (Hill 1948).
(*d*) The ' bare ' wire electrode (Montgomery and Horwitz 1950, Inch 1955, Cater *et al.* 1957).
(*e*) Needle micro-electrode (Cater and Silver 1961).
(*f*) The functional end of a macro-membrane-electrode (Clark 1956).
(*g*) Micro-membrane electrode (Silver 1965).

Figure 3.5: Diagram of various types of polarographic electrodes. Of interest is the "Clark" type electrode, (f), named after its inventor Leyland Clark [29], where the anode and cathode are encased together in the same electrode covered with an oxygen permeable membrane to try and ensure that only oxygen is diffusing into the probe. Clark type electrodes are usually the ones used in medical and biological applications. Figure taken from [31].

blood gas contents, disagrees and claims that "it is not unusual to find differences of [more than] 20%" between different models of polarographic probe when measuring the amount of oxygen dissolved in blood and that the question of accuracy of measurements is "controversial" . In other words, there is significant debate as to whether a polarographic probe is actually useful in determining oxygen concentration in human tissue (or biological tissue in general).

For medical purposes, a needle micro-electrode is used. Typically, this means that the cathode and anode will be made small enough so that they can be encased together behind an oxygen permeable membrane and put in a hypodermic needle suitable for insertion into biological tissue, known as a "Clark type" electrode [31] (see figure 3.5). For instance, Wong et al [12] describes the needle electrode they used as a gold wire 0.012mm in diameter encased in a steel probe 0.3mm in diameter with a teflon membrane at its tip. Vaupel [15] uses a similar instrument, and also notes that the probe has sharply ground tips, and that the gold wire cathode is contained with a recess. As discussed in the mathematical modeling section (3.2), the small size of the probe ensures that the current reaches its steady state quickly: both Wong and Vaupel note that their probes have quick "response times", and Vaupel in particular claims that his had a response time of less than 500ms.

To measure the oxygen tension in a human tumour, the location of the tumour must first be identified. Then under anaesthesia, the needle probe is directly inserted into the predetermined area. Next, a computer automatically advances the probe into the tumour in steps of 1mm forwards and 0.3mm backwards, in what is known as the "pilgrim step" movement. The backward step is to try to alleviate any extra pressure the probe may have caused before taking the oxygen partial pressure readings. In total 20-30 measurements are usually made, and the probe is then retracted and possibly inserted into another location for another track of measurements. Wong [12] claims that it is best to take about 5 tracks of 20-30 measurements each, and that track position does not matter too much, though it is obvious that we should try and spread out the locations of the measurements.

Interpretation of Measurements

To classify the patients based on oxygen tension readings, many measures have been proposed, the one used most often is called hypoxic proportion and it is the one used by Fyles et al at PMH [11]. Hypoxic proportion is defined as the percent-

age of readings under a certain value, usually 5mmHg (abbreviated as HP5), and medical researchers will then use this value instead of the actual oxygen tension measurements to quantify how hypoxic a tumour is. HP5 values usually have a strong correlation to the median oxygen tension measurement [11], [12].

Stadler et al [14] proposes that we multiply the HP5 by the volume of the tumour to get a measure called hypoxic subvolume. He argues that the hypoxic subvolume, which gives us an indication of the number of hypoxic cells, is a better prognostic indicator because the more hypoxic cells there are, the more difficult it is to eliminate them using radiotherapy, and the more likely that there will be therapy resistant mutations.

In an earlier study by Vaupel et al (1991), as opposed to Fyles et al (2002) and Stadler et al (1999), using needle electrodes in breast tumours [15], the actual distribution of oxygen partial pressures were analyzed and compared between normal and malignant tissue, with the median PO2 being the simplest and clearest indicator of tumour hypoxia.

3.3.3 Why Partial Pressure?

One question that arises from the mathematical discussion is: Why do we want the final answer to be in the form of partial pressures? Why apply Henry's law at all? It certainly doesn't make sense to change concentrations to partial pressures in the chemistry applications. The answer seems to come from one of the first uses of polarographic sensors in humans: blood gas analysis. In blood gas analysis, we wish to not only measure the oxygen content in blood, but also in inspired and expired gases, because we want to determine whether machines such as artificial ventilators are working, or whether gaseous anaesthetic agents are having an effect. So it makes sense to talk about partial pressures because we measure the amount of gas in respirators and anaesthetic agents in terms of partial pressures [29].

The idea is that Clark type polarographic probes can actually be used to measure the oxygen partial pressure of gases because there is electrolyte within the sensor itself, which means that oxygen in gaseous form can diffuse into the sensor, dissolve, and become electrolyzed as before. In this case, and similarly with the measurement of other gases, such as anaesthetic agents in inspired and expired gases, we have to

convert back to partial pressures otherwise it wouldn't make sense. Furthermore, despite the fact that most oxygen in blood is chemically bound to haemoglobin and not actually dissolved, Hahn [29] claims the partial pressure of oxygen in blood is an important indication of lung function, and that the partial pressure of inhaled anaesthetic agents determines how much is able to pass through the blood brain barrier .

The point here is that it seems to have become the convention to measure tissue oxygenation in terms of partial pressures so that we can compare the amount of oxygen in tissue to other measurements, such as the amount of oxygen they are inhaling. Indeed, medical researchers such as Fyles [11], Wong [12], and Vaupel [15], all make sure that their patients are breathing in air with a fixed oxygen concentration in an effort to control factors that may affect tumour oxygenation.

3.3.4 Calibration

According to Wong et al [12] at PMH, calibration is performed by placing the probe in sterile 0.9% NaCl solution"though which sterile pure nitrogen and room air flowed alternately". Vaupel [15] describes a similar procedure of immersing the probe in a sterile saline solution, and then calibrating by using pure nitrogen and air. Hahn [29] also uses two liquids with known PO2 concentrations for calibration, with the condition that one of them must have zero PO2.

In all cases, a two point calibration procedure is used assuming that the current is linearly related to oxygen tension as predicted by our diffusion models, allowing the determination of the slope and intercept of the curve so that we can convert current to partial pressure. Hahn states that one of the calibration fluids must be at zero PO2 (both Wong and Vaupel follow this convention when they use pure nitrogen) because he claims that there is a "quiescent current" present in the instrument that is comparable to the output generated by the redox reaction, so he wants to set the zero explicitly in order to adjust the readings appropriately.

As outlined in the section on the membrane covered electrode, Albanese [28] argues that this type of calibration procedure is not adequate as we must account for the different tissue properties. Fatt [24] concurs and claims that such calibration procedures will, at best, give a relative value of oxygen partial pressure. The

natural question, then, that one might ask here is whether the absolute values of the measurements are required, or whether relative changes are sufficient? Silver [31] asserts that relative changes are sufficient for some applications, citing specifically that relative changes are useful in determining the effectiveness of agents designed to raise or lower oxygen tension in tumours in preparation for radiotherapy.

3.3.5 Potential Issues

Several problems with the theory and use of polarographic probes have arisen during the course of the preceding discussion. I wish to reiterate them and raise a couple more here:

1. Does Henry's Law hold in tissue? This is a question that may or may not be important depending on whether it is crucial for us to actually determine the partial pressure. If it is, we may need to revise our models and use a more complex form of Henry's law since the simple form, $C = kP$, that is used only holds under certain ideal conditions (further details on other forms of Henry's Law can be found in Appendix A).

2. When solving for the relationship between pressure and current we made several simplifications, such as the assumption of spherical symmetry of the probe, and the uniform initial distribution. More complicated geometries have been used in the literature, (see [24], [29], [30]) but the current always ends up linearly related to pressure when we assume the simple initial condition.

3. What is diffusing? Physically, it doesn't make sense to say pressure is diffusing, or chemical potential is diffusing, so the quantity in the diffusion equation must be concentration. The fact that we write our diffusion equation with pressure as the variable is misleading, and is only a mathematical expression that holds because we assume the simple form of Henry's Law holds. Mancy's claim that we should correct for activity in higher salt concentrations is not supported by any other researcher.

4. Calibration issues: We have discussed at length the problem of calibrating the probe correctly in tissue. In the next chapter we will see clearly that calibration must be an issue, as there are impossible negative partial pressures in our data. Only Stüben [16] makes any mention of negative values in his results, but he disregards them, and assumes that the readings are in the range 0-2.5mmHg. Other medical researchers such as Wong [12] and Fyles

[11] claim that results are "generally reproducible between different centres" [11] but do not explain what they mean.

5. It is not hard to imagine that inserting a needle into tissue will damage it and potentially affect the polarographic measurements. Toma-daşu [20], [21], briefly mentions the problem in her papers that numerically simulate the polarographic probe, and as discussed previously, doctors use a special movement pattern to try and minimize the damage caused by the probe, but little else has been done. Perhaps of more importance is whether tissue damage caused by the probe affects treatment outcome, and a study by Stüben [16] found that there was no difference between rats that had three needle tracks and rats with eight needle tracks subsequently treated with radiotherapy.

6. Hahn [29] points out that for blood, much of the oxygen is chemically bound to haemoglobin and not dissolved, so for a better indication of oxygen content in tissue we must also consider other factors such as haemoglobin concentration and the affinity of haemoglobin for oxygen .

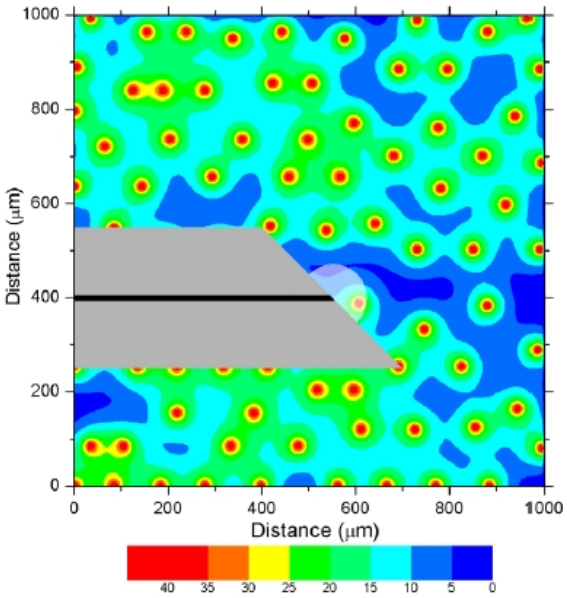


Figure 1. A schematic drawing which shows the relative size of the measurement volume of the polarographic electrode and the spatial scale of variation of the oxygen tension in the tissue.

Figure 3.6: A diagram showing the difference in size between the needle electrode (300μm) and a typical cell (7-15μm). The problem here is that, ultimately, we would like to measure the oxygen content of each individual cell because it is the amount oxygen within the cell that determines whether radiation therapy will be effective. Clearly, the probe is much too large to measure each individual cell, and what it is probably doing is actually measuring the average oxygen content of the cells near the probe. Toma-daşu, [20], [21], has run some computer simulations to show how this effect might come about. Figure from [20]

3.4 Summary

- Oxygen diffuses into the cathode because of the concentration gradient created by the consumption of oxygen at the surface of the cathode: $\frac{\partial C}{\partial t} = D\nabla^2 C$.
- We assume Henry's Law, $C = kP$, holds to get an equation for pressure: $\frac{\partial P}{\partial t} = D\nabla^2 P$.
- An oxygen reduction reaction then occurs at the cathode
  $O_2 + 2H_2O + 4e^- = 4OH^-$
  , that produces a current proportional to the amount of oxygen present $I = nFAf$.
- The flux f is given by Fick's law $j = -Dk(\partial P/\partial x)$, and we set f equal to either $-j$ or $|j|$ to obtain a positive current.
- Modern needle electrodes reach the steady state current in less than 500ms.
- The Cathode (typically gold) is usually about $12\mu m$ in diameter, incased in a needle about $300\mu m$ in diameter, and covered by a membrane (usually teflon).
- The Anode, usually silver-silver chloride, $Ag{+}Cl^- = AgCl{+}e^-$, is a controlled reaction designed to fix the voltage in the cathode so that it stays in the desirable "polarized region".
- To quantify the amount of hypoxia, a measure, HP5, the percentage of readings less than 5mmHg, is often used instead of the actual measurements.

Chapter 4

The Data and Standard (Orthodox) Techniques For Analyzing It

As I mentioned briefly in the introduction, I will be analyzing data from a study conducted the PMH group [11]. The main aim of their study was to consider the effects of hypoxia on patients who had just been diagnosed with malignant cervical cancer and for whom radiation therapy was the only prescribed treatment. The plan was that all of these patients would receive similar dosages and fractionation (45-50 Gy in 1.8-2 Gy daily fractions with 16-25 MeV photons) in similar positions around the pelvis using the same techniques to mimimize side effects (four-field box technique). Once the treatment was administered, the patient was considered disease free, and the idea was to chart the time at which these patients suffered a relapse, or worse, death, and to see whether there was any correlation between the amount of hypoxia and the length of time that a patient managed to stay healthy (known as disease free survival, DFS).

4.1 PO2 in Normal Tissue

In order to understand the Eppendorf data we first need a sense of what the oxygen partial pressure is in normal tissue. Oxygen tension in blood is normally measured using polarographic probes [29], and it is accepted that blood PO2 is about 75-100 mmHg [51] depending on factors such as age and health, so we would expect the oxygen tension in tissue to be roughly comparable. I mention the accepted range

for blood specifically because those measurements were taken after the blood was extracted from the body and hence should be reliable as many of the issues mentioned in the previous chapter have to do with inserting the probe into the tissue.

The only reference to my knowledge that gives values for normal tissue PO2 are the "transcutaneous oxygen monitor" measurements of oxygen tension by Bartlett [52]. Despite its name, it is actually just a Clark-type polarographic probe modified so that it can measure PO2 in tissue simply by attaching it to the skin (hence the name transcutaneous, which means under the skin). Hypoxia is a factor that affects how fast or well wounds heal, and often it is used to help determine whether a patient needs an amputation [52]. From Bartlett's table (figure 4.1) it is reasonable to estimate the oxygen tension in normal tissue to be about 50-70 mmHg; certainly it must be between roughly 30-90 mmHg. Although it is not clear whether readings taken near the skin remain constant further inside the tissue, Bartlett does mention that they were able to use the system to help patients with ulcers, so it would appear that the readings are good for the whole tissue.

4.2 The Data

Two sets of data were received from the PMH group: Oxygen tension data, which is simply a large number of readings from the Eppendorf probe (given in mmHg) and survival or DFS data, which is a list of times that indicate whether the patient had a relapse of cancer after the initial radiation treatment, or whether they were healthy on last examination. In this thesis I will primarily be analyzing the first set of data, though the ultimate goal is to see if there is any correlation between the two.

The Eppendorf data contains the track number, and the position along the track where the measurement was taken (see figure 4.2). There are many anomalous readings, such as negative values for pressure, extremely high values (>100mmHg) that we would not expect in a tumour (figure 4.3), and sometimes large oscillations between each reading (figure 4.5) that may be due to the extreme conditions in a tumour or may be noise in the measurements. However, most of the readings are in the low PO2 range (0-10 mmHg), which is what we would expect in a tumour (figure 4.4).

As mentioned before, for reasons that aren't entirely clear, researchers prefer

TISSUE OXYGEN TENSION VALUES FOR PROGRESSIVELY INCREASED INSPIRED PO_2

Ambient pressure (atm abs)/ Breathing media	1.0 AIR	1.0 O_2	2.0 O_2	2.4 O_2	3.0 O_2
REPRESENTATIVE TISSUE OXYGEN TENSION VALUES, mmHg					
Ambient PO_2, mmHg	159	760	1,520	1,824	2,280
Transcutaneous PO_2 (1)	69 ± 6	440 ± 95		1,350 ± 220	
Transcutaneous PO_2 - chest (2)	67 ± 12	450 ± 54		1,312 ± 112	
Transcutaneous PO_2 - calf, male (2)	49 ± 14	281 ± 78		1,027 ± 164	
Transcutaneous PO_2 - calf, female (2)	59 ± 12	367 ± 59		1,174 ± 127	
Transcutaneous PO_2 - midfoot (2)	63 ± 13	280 ± 82		919 ± 214	
Transcutaneous PO_2 - limb (3)	49	325	696		

1 - TRANSCUTANEOUS O_2 DATA FROM SHEFFIELD (1998)
2 - TRANSCUTANEOUS O_2 DATA FROM DOOLEY (1997)
3 - TRANSCUTANEOUS O_2 DATA FROM HART (1990)

(CHART FROM PJ SHEFFIELD, 1998)

Figure 4.1: Table of PO2 values for normal tissue under different inspired PO2. The measurements are labeled as "transcutaneous" (under the skin) because they were taken using a device that measures PO2 by attaching several probes onto the skin. The point here is that, under normal conditions, oxygen tension in tissue should be about 50-70 mmHg, depending on the particular part of the body. Figure from [52]

not to use the actual readings, but a measure called hypoxic proportion. We will define them more explicitly here for reference later:

- **Hypoxic Proportion** is defined as the percentage of readings under a certain threshold value.

Many papers choose 5mmHg as their threshold [11][17], so it is convenient to abbreviate:

- **HP5** is defined to be the percentage of readings under 5mmHg.

In particular, the PMH group chose 5 mmHg because the median PO2 in their case was about 5 mmHg. However, not all researchers choose 5mmHg; some choose 10 mmHg (HP10) [19], or even 2.5 mmHg (HP2.5) [18]. These three values (2.5, 5,and 10 mmHg) seem to be the standard choices, although there appears to be no reason why one couldn't choose 3 mmHg or 7 mmHg or some other arbitrary value. In

1	Study Number	Track Number	Position Along Track	PO2
2	1	1	1	-1.3
3	1	1	2	15.2
4	1	1	3	9.7
5	1	1	4	30.6
6	1	1	5	37.7
7	1	1	6	19.1
8	1	1	7	8.5
9	1	1	8	9.4
10	1	1	9	8.9
11	1	1	10	21.4
12	1	1	11	23.5

Figure 4.2: Here are some of the oxygen tension data. There are over 13000 measurements in total. PO2 values are given in mmHg. Notice the impossible negative value for partial pressure right at the beginning.

11756	P02	2	7	70.1
11757	P02	2	8	87.4
11758	P02	2	9	103
11759	P02	2	10	105
11760	P02	2	11	103
11761	P02	2	12	104
11762	P02	2	13	106
11763	P02	2	14	114
11764	P02	2	15	115
11765	P02	2	16	120
11766	P02	2	17	114
11767	P02	2	18	108
11768	P02	2	19	105
11769	P02	2	20	108
11770	P02	2	21	113
11771	P02	2	22	109
11772	P02	2	23	105
11773	P02	2	24	112
11774	P02	2	25	108
11775	P02	2	26	100
11776	P02	2	27	22.1
11777	P02	2	28	123
11778	P02	2	29	120
11779	P02	2	30	101
11780	P02	2	31	163
11781	P02	2	32	211
11782	P03	1	1	7.6
11783	P03	1	2	11.5
11784	P03	1	3	16.1
11785	P03	1	4	19
11786	P03	1	5	20.5

248	3	1	21	46.6
249	3	1	22	46.7
250	3	1	23	46.6
251	3	2	1	-0.5
252	3	2	2	-0.5
253	3	2	3	-0.6
254	3	2	4	-0.6
255	3	2	5	-0.6
256	3	2	6	-0.5
257	3	2	7	-0.5
258	3	2	8	-0.6
259	3	2	9	-0.5
260	3	2	10	-0.6
261	3	2	11	-0.5
262	3	2	12	-0.5
263	3	2	13	-0.5
264	3	2	14	-0.5
265	4	1	1	0
266	4	1	2	3.9
267	4	1	3	20.9
268	4	1	4	29.9

Figure 4.3: Here are some anomalous values. As mentioned, normal PO2 in human tissue is about 50-70 mmHg so these values are abnormally high even for normal tissue (higher than even PO2 in air, which is 159 mmHg according to the table on the previous page), or impossibly negative.

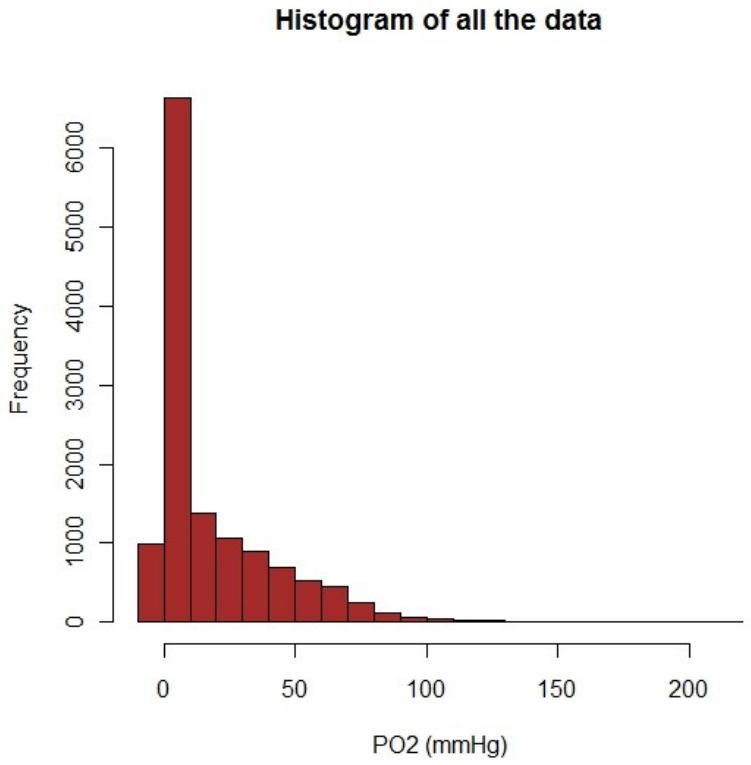


Figure 4.4: Histogram of all the PO2 data. There were 105 patients, each with about 3-4 tracks of data, and 20-30 measurements per track, amounting to over 13000 measurements. The maximum value was 212.8mmHg, and the minimum value was -4.3mmHg. From the histogram we can clearly see that most of the data is between 0-10mmHg and that a surprisingly large portion is negative.

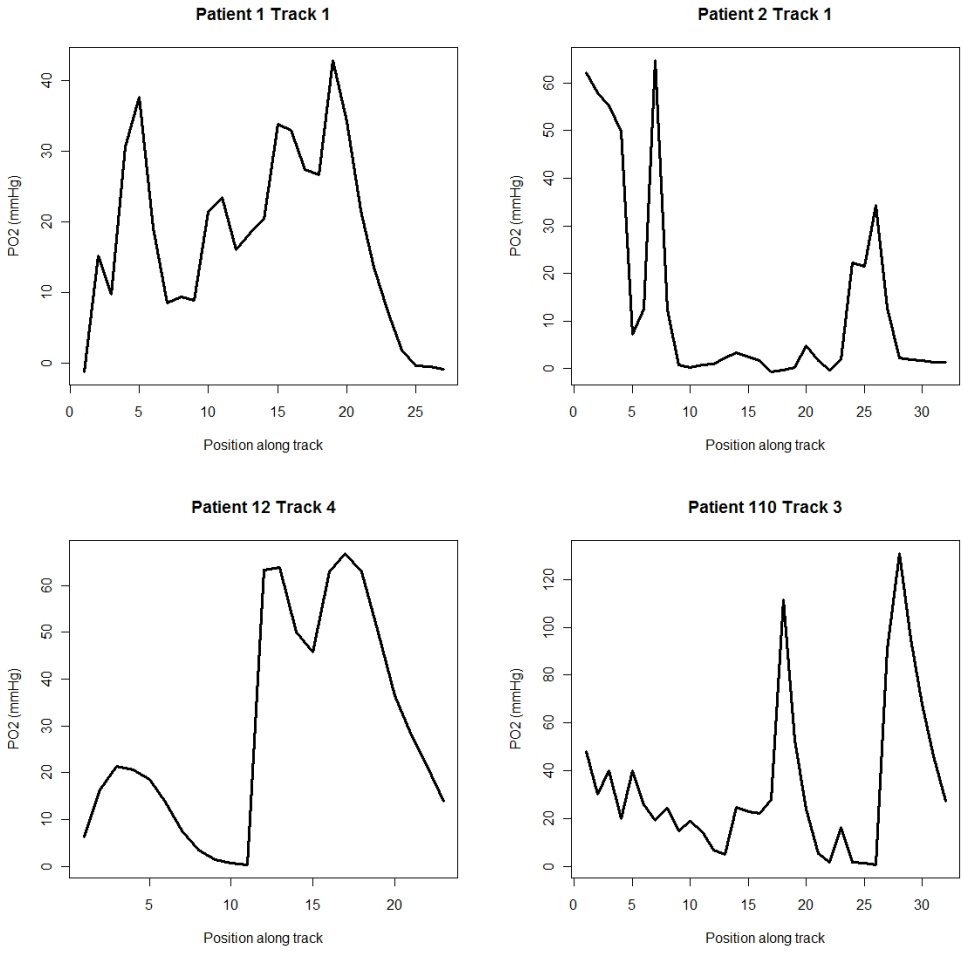


Figure 4.5: Some plots of oxygen tension versus position. Note the large oscillations between readings physically millimeters apart that would indicate some kind of noise or extreme conditions.

most cases, the choice seems to depend on which will give the lower p-value in the log-rank test. I have chosen to report HP5 in subsequent chapters as a decimal (e.g. 0.56) as opposed to a percentage (56%) for convenience in typesetting.

4.3 Correlation between HP5 and Median PO2

Fyles et al [11] find a strong correlation between HP5 and median PO2, based on the calculation of the Spearman coefficient of rank correlation, which is given by

$$R = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}. \tag{4.1}$$

Briefly, to calculate the Spearman coefficient, for a measured set of paired data (x_i, y_i) (in our case the HP5 and median PO2 for each patient), the first step is to rank each x_i and y_i in either ascending or descending order (doesn't matter which as long as you do the same to both) to get a new set of ranked data (X_i, Y_i). Then, this set of paired ranked data are used as an approximation of the true correlation coefficient ρ, instead of the actual data, using the formula for R above - see Appendix B for a full derivation of equation 4.1.

The Spearman correlation coefficient, with $-1 \leq R \leq 1$, gives us a sense of whether the pairs (x_i, y_i) vary together, that is, whether if the increase in one will have any effect on the increase or decrease of the other. Fyles et al obtain $R = -0.91$ for HP5 and median PO2 and, using what should be the same data, I obtain $R = -0.92$. Thus, this means that it is very likely that as the median PO2 increases HP5 will decrease.

For comparison, consider the usual correlation coefficient in standard orthodox statistics [47]

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}}, \tag{4.2}$$

which is often written as

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}, \tag{4.3}$$

HP5	Median PO2		Hp5 Ranked	PO2 Ranked	d	d^2
0.520	4.7		57	51	6	36
0.438	12.45		47	68	-21	441
0.568	3.2		63	38	25	625
0.468	6.75		52	57	-5	25
0.917	0.6		103	7	96	9216
0.127	31.85		18	91	-73	5329

Figure 4.6: To calculate the Spearman correlation coefficient between two sets of values, we first rank each set in order, calculate the square of the difference in ranks, then plug them into the formula 4.1, with $D_i = X_i - Y_i$.

where $\bar{x}$, $\bar{y}$ are the sample means, and s_x, s_y are the sample standard deviations, given by

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{4.4}$$

and similarly for s_y. Using this measure of correlation we get $r = -0.74$, which still indicates negative correlation, though not as strong. The difference between the two measures of correlation is that for the standard correlation we are testing for linear dependence between the two variables, whereas for Spearman we do not make any assumptions about the form of the population distribution.

In my opinion, real question here is why one even uses HP5 or median PO2 in the first place? There is no real mathematical or biological justification for either choice in any of the medical literature (e.g. [11][12][14][15][16][17][18][19]) and we will see shortly that it is not completely clear whether there is any correlation between HP5 and survival.

4.4 Survival Data and its Analysis

The survival data is simply a list of two numbers for each patient (see figure 4.7). "Event" indicates whether the patient has suffered a relapse or death (labeled as 1), or whether he or she was alive and healthy during the last checkup (labeled as 0). The column labeled "Time" is known as the study time, and it is the time from when that particular patient entered the study to the time of relapse or death, or the time of last checkup.

1	Study Number	Event	Time
2	1	1	0.939
3	2	0	7.838
4	3	0	9.706
5	4	0	7.313
6	6	1	0.589
7	7	1	0.003
8	8	1	9.177

Figure 4.7: Here is some of the survival data. Event indicates whether the patient has suffered a death or relapse (1) or whether he was disease free at time of last checkup (0). Time is in years.

Since the event that we want to consider is relapse of the cancer or death due to the cancer, the survival data will often be incomplete because not all patients will have encountered that event at the time of analysis. For example, a patient could have stayed healthy and so we do not know when (or if) he or she will have a relapse, or, another patient might move out of the country and not return for regular checkups so we lose contact with them, or, yet another patient might have died in a car crash and not because of the cancer, so again, the relapse time is uncertain. In any case where we do not have information as to when exactly the patient suffered a relapse all we know is that the time to the event is *at least* as much as what we have recorded. In survival analysis this kind of incomplete data is called right censored because if we think of time as the x-axis, then we do not know what happens past a certain point to the right of the graph.

Both Collet[45] and Pintilie [46] mention that it is important that the mechanism that caused the patient to be censored must be independent of the survival time t, called "non-informative censoring", so that a censored patient is just as representative of a patient who survived up to time t as any other patient. For instance, a patient who died because of catastrophic liver failure due to the cancer metastasizing would be "informative censoring" because the cause of death is related to the event we are interested in. Whereas dying in a car crash due to a drunken driver would be non-informative. We will assume, or make sure, that all the censoring in our data is non-informative to simplify the analysis.

In survival analysis we wish to estimate the survival probability, or **survival**

function, of the patients over time. Mathematically, following Collet [45], let T be a random variable that represents the survival time of the patient. We define the survival function, $S(t)$, as the probability that the patient will have survived at least up to time t:

$$S(t) = P(T \geq t) = 1 - F(t), \tag{4.5}$$

where $F(t)$ is cumulative probability distribution function of T

$$F(t) = P(T < t) = \int_0^t f(u)du, \tag{4.6}$$

and $f(u)$ is the probability density function for T. In many cases, researchers prefer to express their result in terms of the hazard function, which is the instantaneous death (or relapse) rate at time t, provided he or she has survived up to that time:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}. \tag{4.7}$$

In other words, $h(t)\delta t$ is the probability that a patient dies (or suffers a relapse) in the interval $(t, t + \delta t)$ conditioned on that person having survived to time t. For instance, if we measure time in days, then $h(t)$ is approximately the time the patient will die the next day if he or she were alive on day t. Using the product rule, we can rewrite the hazard function in the following way:

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}, \\ &= \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t)}{\delta t P(T \geq t)} \right\}, \\ &= \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}, \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

Comparing this to 4.5 and 4.6 we get

$$h(t) = -\frac{d}{dt}\{log\ S(t)\}. \tag{4.8}$$

If we define the cumulative hazard function $H(t)$ as

$$H(t) = \int_0^t h(u)du, \tag{4.9}$$

then the survival function can be written as

$$S(t) = exp\{-H(t)\}. \tag{4.10}$$

Depending on the situation, one might want to express the probability of survival in terms of any one of these functions. In any case, the idea is that if we can estimate one of $S(t)$, $h(t)$, or $H(t)$ from our data then we can derive the others as desired.

For some insight on how to estimate survival, first consider the simple case where there are no censored patients. In that case, an obvious choice would be the **empirical estimate**:

$$\hat{S}(t) = \frac{\text{Number of individuals with survival times} \geq t}{\text{Number of individuals in the data set}}. \tag{4.11}$$

Notice that the estimated probability of survival is exactly 1 before any patients die, and is exactly 0 after all patients have died. We assume $\hat{S}(t)$ to be constant between deaths, so it is simply a step function that drops down every time there is a death.

Example. *Following Collet [45], suppose we have the following list of survival times for something that eventually fails, say the time at which light bulbs burn out:*

$$11 \quad 13 \quad 13 \quad 13 \quad 13 \quad 13 \quad 14 \quad 14 \quad 15 \quad 15 \quad 17 \tag{4.12}$$

The empirical survival function will be constant between the failure times 11, 13, 14, 15, and 17, and the value in those intervals are respectively 0.909, 0.455, 0.273, and 0.091, attained by using equation 4.11. See figure 4.8.

For censored data, the **Kaplan-Meier estimator** is the most commonly used method for estimating the survival function. It was first proposed by Kaplan and Meier in 1958 [48] and actually reduces to the empirical estimate in the case where there are no censored data.

Suppose that we have survival times for n patients, $t_1, t_2, ..., t_n$. Some of these may be right censored, so we suppose that we have r unique death times (i.e. more than one patient may die at that time), with $r \leq n$, and order those death times in ascending order $t_{(1)} < t_{(2)} < ... < t_{(r)}$. Let n_j, $j = 1, 2, ..., r$, be the number of individuals who survive up until time $t_{(j)}$ (known as the number at risk at $t_{(j)}$), and d_j be the number who die at this time. Now consider the time interval $[t_{(j)}, t_{(j+1)})$. As before, we assume that the probability of survival is constant between events, so from the way we have constructed the intervals, the probability of survival in this interval is the probability of surviving the event at $t_{(j)}$, and that is estimated by

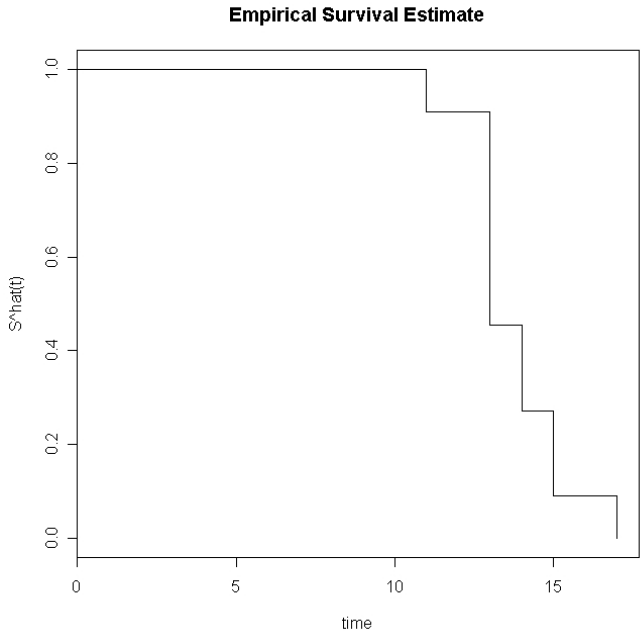


Figure 4.8: A plot of the estimated survival function from example 4.4.

$(n_j - d_j)/n_j$. Assuming that deaths occur independently, we get that the estimated survival function is given by

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{4.13}$$

for $t_{(k)} \leq t < t_{(k+1)}$, with $k = 1, 2, ..., r$, and $t_{(r+1)}$ taken to be ∞. Notice that $\hat{S}(t) = 1$ for $t < t_{(1)}$ as before, and $\hat{S}(t) = 0$ for $t > t_{(r)}$ if $t_{(r)}$ is an actual death time and not a censored time. Also when there are no censored patients, n_j would be equal to $n_{j-1} - d_{j-1}$ for all j, thus:

$$\begin{aligned} \hat{S}(t) &= \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \\ &= \frac{n_k - d_k}{n_k} \cdot \frac{n_{k-1} - d_{k-1}}{n_{k-1}} \cdots \frac{n_1 - d_1}{n_1}, \\ &= \frac{n_k - d_k}{n1}. \end{aligned}$$

Since $n_1 = n$ and $n_k - d_k$ would be the number of survivors at time t, this is exactly the empirical survival estimate 4.11. In other words, the Kaplan-Meier estimate is simply a more sophisticated version of the empirical estimate that takes into

account right censored patients.

The preceding was a somewhat heuristic justification for the Kaplan-Meier estimator modified from Collett [45] and it can actually be derived using maximum likelihood as follows from Pintilie [46]. In the discrete case where the random variable for survival time T can only take on discrete values $t_1, t_2, ...$, with $0 \leq t_1 < t_2 < ...$, the survival function is given by

$$S(t) = P(T > t) = \sum_{t_j > t} P(T = t_j). \tag{4.14}$$

The hazard function is

$$\begin{aligned} h(t_j) = P(T = t_j | T > t_{j-1}) &= \frac{P(T = t_j)}{S(t_{j-1})}, \\ &= 1 - \frac{S(t_j)}{S(t_{j-1})}, \end{aligned} \tag{4.15}$$

and so induction yields

$$S(t) = \prod_{t_j \leq t} \left[1 - h(t_j)\right]. \tag{4.16}$$

Now let m_j be the number of individuals censored in the interval $[t_{(j-1)}, t_{(j)})$. The likelihood function is

$$L = \prod_{j=1}^{r} \left[S(t_{(j-1)}) - S(t_{(j)})\right]^{d_j} S(t_{(j-1)})^{m_j}, \tag{4.17}$$

then, substituting $S(t)$ for the hazard $h(t)$ using equation 4.16, and using the short

hand $h_j = h(t_j)$, provides the following simplification:

$$\begin{aligned}
L &= \prod_{j=1}^{r} \left\{ \left[\prod_{i=1}^{j-1}(1-h_i) - \prod_{i=1}^{j}(1-h_i) \right]^{d_j} \left[\prod_{i=1}^{j-1}(1-h_i) \right]^{m_j} \right\}, \\
&= \prod_{j=1}^{r} \left\{ \left[\prod_{i=1}^{j-1}(1-h_i) \right]^{d_j} [1-1+h_j]^{d_j} \left[\prod_{i=1}^{j-1}(1-h_i) \right]^{m_j} \right\}, \\
&= \prod_{j=1}^{r} \left\{ h_j^{d_j} \left[\prod_{i=1}^{j-1}(1-h_i) \right]^{d_j+m_j} \right\}, \\
&= \left\{ \prod_{j=1}^{r} h_j^{d_j} \right\} \left\{ \prod_{j=1}^{r} \prod_{i=1}^{j-1} (1-h_i)^{d_j+m_j} \right\}, \\
&= \left\{ \prod_{j=1}^{r} h_j^{d_j} \right\} \left\{ \prod_{i=1}^{r} \prod_{j=i+1}^{r} (1-h_i)^{d_j+m_j} \right\}, \\
&= \left\{ \prod_{j=1}^{r} h_j^{d_j} \right\} \left\{ \prod_{i=1}^{r} (1-h_i)^{\prod_{j=i+1}^{r}(d_j+m_j)} \right\}, \\
&= \left\{ \prod_{j=1}^{r} h_j^{d_j} \right\} \left\{ \prod_{i=1}^{r} (1-h_i)^{n_i-d_i} \right\}, \\
&= \prod_{j=1}^{r} h_j^{d_j} (1-h_j)^{n_j-d_j}.
\end{aligned}$$

Taking logarithms of both sides and differentiating with respect to h_j gives

$$\frac{\partial log\ L}{\partial h_j} = \frac{d_j}{h_j} - \frac{n_j - d_j}{1-h_j}. \tag{4.18}$$

So the maximum occurs when

$$\hat{h}_j = \frac{d_j}{n_j}, \tag{4.19}$$

and hence the maximum likelihood estimate of the survival function $S(t)$ occurs when this value of $h(t)$ is plugged into equation 4.16 to get

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j} \right) = \prod_{t_j \le t} \left(\frac{n_j - d_j}{n_j} \right), \tag{4.20}$$

which is exactly equation 4.13 but with slightly different notation. For information on how to calculate and derive quantities such as the variance and 95% confidence interval for the Kaplan-Meier estimator, the reader is referred to Collett [45] or Pintilie [46]. The results of applying the Kaplan-Meier estimator to the data from

```
> fit<-survfit(Surv(Time,Event),data=survdata)
> summary(fit)
Call: survfit(formula = Surv(Time, Event), data = survdata)

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.003    106      17    0.840  0.0356        0.773        0.912
 0.460     89       1    0.830  0.0365        0.762        0.905
 0.512     88       1    0.821  0.0373        0.751        0.897
 0.589     87       1    0.811  0.0380        0.740        0.889
 0.690     86       1    0.802  0.0387        0.729        0.881
 0.706     85       1    0.792  0.0394        0.719        0.874
 0.709     84       1    0.783  0.0400        0.708        0.866
 0.868     83       1    0.774  0.0406        0.698        0.858
 0.939     81       1    0.764  0.0413        0.687        0.849
 0.991     80       1    0.754  0.0418        0.677        0.841
```

Figure 4.9: This is what R, a commonly used free statistical computation program, gives when calculating the first few Kaplan-Meier estimates. The first entry is peculiar because the data I got had 17 patients all dying or suffering a relapse at t=0.003. This would suggest that these 17 entries are actually left censored, that is, the actual time of event is less than 0.003 but was only discovered then because that is when they had a routine check up. However, we will assume that is not the case. R uses slightly different notation than I have in the text: time are the event times $t_{(j)}$, n.risk is n_j, and n.event is d_j. More information on how the standard error and confidence intervals are calculated can be found in [45] and [46].

the PMH group are shown in figures 4.9 and 4.10.

A further analysis of the survival data was done in order to test whether HP5 has any effect on survival. Fyles et al split the patients into two groups based on their HP5: they classify patients with HP5>.5 as hypoxic (i.e. badly oxygenated) and patients with HP5<.5 as oxic (i.e. well oxygenated), and the survival probability of the two groups were compared. Similar to the choice of 5mmHg as the threshold for hypoxic proportion, the choice of 50% as the threshold for splitting is also arbitrary and is again because the median HP5 is about 50% [11]. The idea is that, based on our discussion on the oxygen fixation hypothesis and how hypoxia adversely affects radiotherapy, we would expect the hypoxic group to fare much worse than the oxic group, and using the Kaplan-Meier estimator and **log-rank test** (reviewed in Appendix C) we would like to see if that is actually the case.

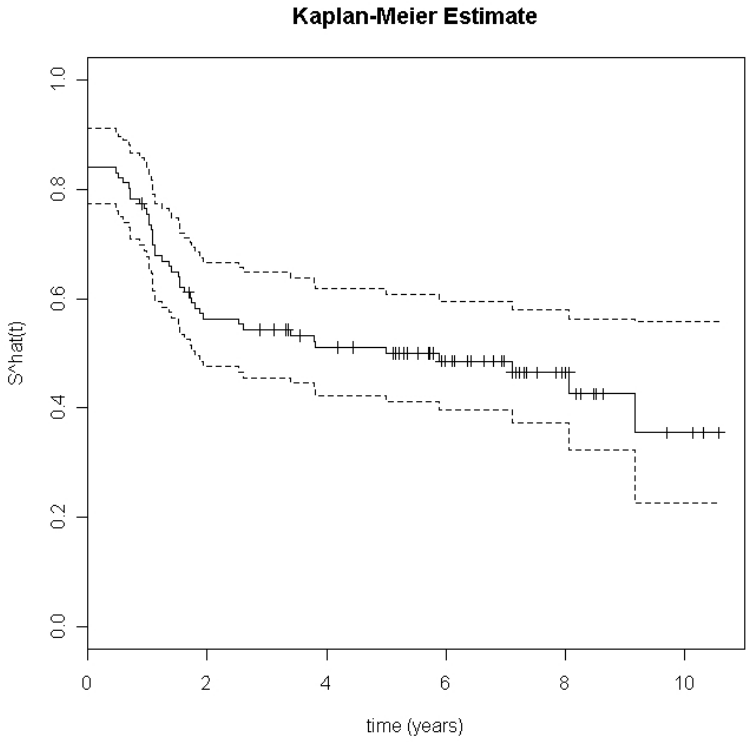


Figure 4.10: This is a Kaplan-Meier plot of the survival data from the PMH group. The solid curve in the middle is the Kaplan-Meier estimate of the survival function, the vertical dashes on the curve indicate places where patients were censored out of the study, and the dotted lines indicate the 95% confidence interval about the Kaplan-Meier estimate. Care must be taken when interpreting survival plots such as these because we have assumed that the survival probability is constant between events. Peto [44] warns us that "any conclusion based on the fine detail of such a graph is likely to be wrong" and that it is only the general shape of the curve that is informative. Despite this warning, these graphs are often used to estimate, say, the "three year survival rate" which from my plot is roughly 54%, in broad agreement with the 53% reported in Fyles et al [11].

To begin, it is often quite informative to simply compute and plot the Kaplan-Meier curves for each group. As we can see in the plot of our data in figure 4.11, the hypoxic group clearly has a much lower survival probability, and from that we can surmise that there is probably some correlation between HP5 and survival. In particular, we might conjecture that patients with high HP5 (i.e. more hypoxic tumours) are less likely to survive.

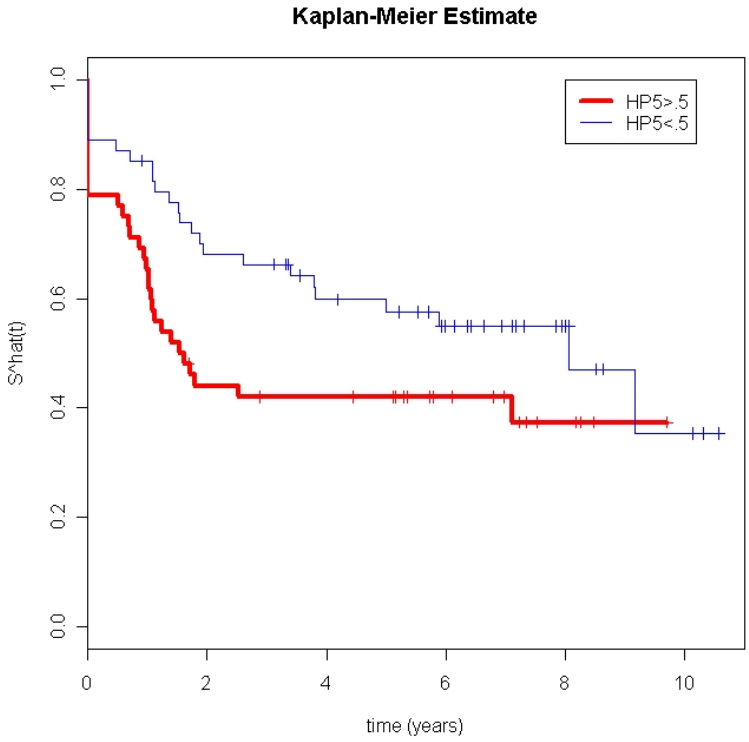


Figure 4.11: A plot of the data stratified by HP5 into two groups. It is fairly clear from this plot that one group has a higher chance of survival, at least for the first 8 or 9 years or so.

Simply making observations from a plot is fairly vague, so we turn to the log-rank test to try to quantify how much worse the patients with hypoxic tumours are doing compared to the oxic group. For the data that I received, I obtained a log-rank p-value of 0.063 when comparing the survival of the oxic group to the survival of the hypoxic group. However Fyles et al [11] report that they obtained p=0.004, so there is a (rather large) discrepancy between our results. The most likely explanation is that the survival data that I received in 2006 had been updated since the original publication in 2001. Indeed, the PMH group themselves

published a revised paper in 2006 [13] in which their log-rank p-value for HP5 was now 0.05, which is in much closer agreement with my calculation.

4.5 Some Critical Comments

The discrepancy between the Fyles et al's initial result and the other two results is important because in orthodox statistics it is conventional to consider a p-value of less than 0.05 to be "statistically significant". That is, we normally require a p-value of (much) less than 0.05 to reject the null hypothesis, and the later results do not allow us to do so, seemingly indicating that the updated data undermines the result of their earlier study. However, the choice of 0.05 for statistical significance is once again completely arbitrary, so we might still find that the null hypothesis is not plausible given the higher, but still very low, p-values. Further, not being able to reject the null hypothesis may seem like a problem at first glance, but in reality, it is only marginally less informative than if we had been able to reject it.

```
> survdiff(Surv(Time,Event)~HP5,data=survdata)
Call:
survdiff(formula = Surv(Time, Event) ~ HP5, data = survdata)

       N Observed Expected (O-E)^2/E (O-E)^2/V
HP5=0 52       31     24.3      1.85      3.46
HP5=1 54       25     31.7      1.42      3.46

 Chisq= 3.5  on 1 degrees of freedom, p= 0.063
```

Figure 4.12: Here is what R outputs when computing the log-rank test. As mentioned earlier, all we need are the total number of observed deaths and the total number of expected deaths in order to compute the log-rank statistic, shown in the column on the far right.

The issue here is whether "statistically significant" is the same as practical significance, and whether we require the result of the log-rank test to validate our claim that patients with more hypoxic tumours are less likely to survive. Armstrong [22][23] argues that the answer is no to both those questions. He claims that there is a large body of evidence to show that, in many cases, significance tests such as the log-rank test are misleading and misinterpreted [22], that matters

which are statistically significant are often not physically relevant, and conversely that factors that are practically important are often discounted by significance testing. Moreover, he asserts that there is absolutely no empirical evidence to support that significance testing actually helps researchers under any circumstance at all and has in no way advanced scientific understanding in any area [22][23]. Thus, Armstrong calls for scientific journals, especially those in the social sciences, to stop requiring that researchers include results of significance testing as (erroneous) evidence to support their claims.

The main problem with significance testing is the null hypothesis. The null hypothesis is often chosen for its simplicity, and Armstrong [22] remarks that some researchers mockingly call it the "nil hypothesis" because often a rejection of such a simple hypothesis tells us nothing new. In our case, the rejection of the null hypothesis simply tells us that the survival probability of the two groups are different, not how or why they are different, just that they are different, and we knew that already from the Kaplan-Meier curves. Rejection of the null hypothesis only eliminates one possibility out of a large number of potential explanations and does not shed any light on any of these other explanations.

Chapter 5

Bayesian Data Analysis - The Basics

Before we consider in depth what it means to analyze data from a Bayesian perspective, perhaps it is best to consider a concrete example. With the widespread use of the internet and email these days, much of the email that we receive is junk, or spam, and we depend on spam filters to eliminate these potentially harmful (virus ridden) emails from our inboxes. It may come as a surprise that spam filters actually make use of probability theory to determine what is and isn't worth our time. To determine whether an email is spam, we need to consider its contents. For instance, how do we know if an email is spam if it contains the word "free"? The email in question could very well be spam, trying to trick its recipients by offering free trips or gifts and so on, or it may be an email from a colleague about "free form solutions" and "free energy systems", or it may even be a legitimate email from a friend who actually got a gift and is telling you how he got it. How do you know? A human would be able to deduce, to a great degree of certainty, whether the email was spam simply by reading it. He or she would be able to tell by noting who the sender was, how the email was written, what tone of voice was used, and so on, but a computer doesn't posses the cognitive ability to distinguish spam from legitimate emails, it doesn't understand how to decipher meaning from prose (since humans don't exactly know how their brains do it either), so the best a computer program can do, based on prior user and programmer input on what is and ins't spam, is offer a probability.

To figure out the probability that an email is spam given that the word "free" is used, we need to appeal to Bayes' theorem to reverse the dependence and get the

probability that we want in terms of quantities that we can obtain:

$$P(\text{An email is spam}|\text{free}) = \frac{P(\text{free}|\text{An email is spam})P(\text{An email is spam})}{P(\text{free})}. \tag{5.1}$$

The probability on the left hand side is the one that we want to calculate and it is normally referred to as the *posterior probability*. The probabilities on the right are all ones that we can estimate by gathering data:

- $P(\text{free}|\text{An email is spam})$ is the probability that a spam email contains the word "free". This probability can be estimated initially by collecting a large number of known spam emails and noting whether "free' appears in them. It is then customizable for each user by flagging certain emails as spam, or by telling the filter that a certain email it thought was spam is actually not. This ability for the filter to learn as new information becomes available is one of the key advantages of the Bayesian formulation. In general, we call this probability the *likelihood function*.

- $P(\text{An email is spam})$ is the probability that you would get a spam email in general, which we can estimate by collecting data on a large number of emails, and is independent of the user since we would expect, on average, for each person to have the same probability of receiving spam. That may not be the case, the user might be more susceptible to spam because he or she published their email on a website to attract more attention to their online store say, and so in reality this probability is conditional on the user and possibly other prior information. That is, it should really be denoted as $P(\text{An email is spam}|\text{user})$. This probability is normally called the *prior probability*.

- $P(\text{free})$ is the probability that the word "free" will appear in any given email (written in English). Again, we can estimate this by looking at lots of emails since this probability is also independent of the recipient. We call this probability the *evidence*.

If we wanted to consider more words, let's denote them as some set $\{words\}$, we would simply extend the previous formulation to consider joint probabilities instead:

$$P(\text{An email is spam}|\{words\}) = \frac{P(\{words\}|\text{An email is spam})P(\text{An email is spam})}{P(\{words\})} \tag{5.2}$$

A more thorough and entertaining discussion on Bayesian spam filtering can be found online in an article by Paul Graham [7]. He proposes methods for dealing with more complicated situations such as when there is capitalization (FREE) and when there is punctuation (FREE! vs FREE!!!! etc), and notes that a filter that only deals with the appearance of individual words, and not more complex structures such as where the word appears in a sentence and so on, is adequate for catching most spam. The only problems he has are with emails that are purportedly from members of the opposite gender soliciting discrete affairs that do not use words which commonly appear in spam. What's most interesting is that Graham claims that if the word "free" appears in the subject line of an email, then the probability that it is spam is an astonishing 99.99%!

In essence, what we have done here is to tell the computer how to make an educated guess as to whether a given email is spam or not, given the appearance of certain words, sentence structures, attachments and so on. In other words, we told it how to make a *best guess.* In general, the process of making a guess based on available data and previous information is known as induction or logical inference. Some might call the process prediction, but we will see that they are essentially the same thing. Bayesian inference methods quantify this process by taking advantage of probability theory and Bayes' theorem.

Key questions that arise from this brief discussion:

- Why probability theory?
- What does Bayes' theorem mean?
- How do you prescribe the prior, likelihood function, and evidence?

5.1 Induction and Plausible Reasoning

The idea of induction or plausible reasoning is one that we encounter everyday. Imagine that there are ominous dark clouds in the sky, what do you do? Do you bring an umbrella? Do you stay home? How much faith do you have in the weather report? Will there be a traffic jam on the highway? How do you make those decisions? The best we can do take into account previous experiences (e.g. snow storms usually occur around this time of year), along with the current information (the dark clouds and weather reports), and make a guess as to what will happen next because we are not God, and we do not have any means of knowing for certain whether it will rain or not. In the words of E.T. Jaynes [1], we live in a world of "incomplete information", that is, we only have access to what we can observe (such as the dark clouds) and not the underlying processes. We could develop a complicated model with millions of variables, compute the equations on a super-computer, and find that our model does a good job of predicting the weather, but we will never be 100% sure that we are correct, and just one unexpected storm will throw our equations into doubt. In other words, we can only make reasonable inferences, educated guesses if you will, and not deductions about almost everything in life.

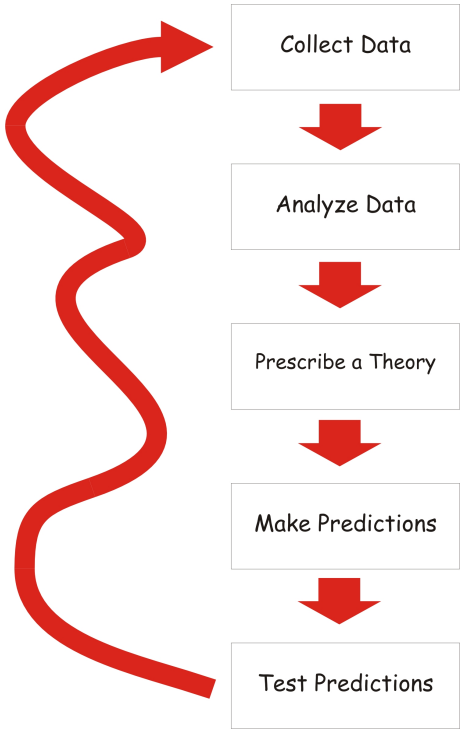


Figure 5.1: The scientific method.

Further, consider for a moment what we would normally describe as the scientific method. For a physical phenomenon that we are interested in, we follow the scheme summarized in figure 5.1 to conduct our experiments. Thus one might argue that analyzing data and performing logical inference based on the data is the key to doing real science, that discovering hidden patterns and information within the data will lead to new scientific breakthroughs, and hence that developing powerful and consistent ways of performing data analysis and logical inference is important and meaningful.

In general, plausible reasoning is the way through which we perform logical inference. If we see dark clouds in the sky, it seems *more plausible* that it will rain; it might not, but one would be *more inclined to believe* that it will. How do we quantify this mathematically?

5.2 Probability Theory - Rules for Consistent Reasoning

It turns out that standard probability theory is precisely what we need to describe the process of logical inference and plausible reasoning.

5.2.1 Degrees of Belief

The first question we need to answer is how do we quantify belief, or rather, *relative degrees of belief*? One thing that humans do is they have some kind of measure of how much we believe something, which we will call our degree of belief. Belief is not black or white - we do not believe something to be true just because someone tells us it is so - most humans are naturally skeptical unless presented with some kind of evidence, and even then we may question the evidence before us. However, belief can be many shades of grey. If Stephen Hawking tells us that black holes exist, we tend to believe him even though we may not even understand what black holes are. On the other hand, if Barack Obama tells us not to worry, and that he will solve the financial crisis, we might not believe him so readily. The point here is that we are capable of different degrees of belief, capable of believing one thing more than another, and as intractable as that may sound, such a measure of belief has to follow certain logical rules.

Going back to the rain example, you would be more likely to believe the weatherman than your butcher on whether it will rain in the next few hours, and likewise more likely to believe your butcher than your infant son on the same matter, so it stands to reason that you would always prefer the forecasts of the weatherman over those of the baby. Such reasoning is *transitive*; that is, if we believe A more than B, and B more than C, then it has to be that we believe A more than C. Otherwise, our logic would be inconsistent and nonsensical. **Real numbers have precisely**

this property, and so we quantify degrees of belief with real numbers, assigning larger numerical values to propositions we believe more.

5.2.2 Sum and Product Rule

Once we accept that degrees of belief can be represented by real numbers, we need to develop rules to specify logical reasoning. According to Sivia [2], Richard Cox in 1946 reasoned that if we specify how much we believe a certain proposition, X, to be true, then we have automatically specified how much we believe it to be false. This piece of simple insight eventually led to the *sum rule* of probability:

$$P(X|I) + P(\overline{X}|I) = 1, \tag{5.3}$$

where $P(X|I)$ is the probability that X is true given some background information I, and $\overline{X}$ denotes the proposition that X is false. At the two extremes, it is convention that we have $0 = P(false)$, and $1 = P(true)$. Cox also argued that if we specify how much we believe Y to be true, and also how much X is true given that Y is true, then we have also specified how much we believe X and Y to be true at the same time, giving the *product rule* of probability:

$$P(X, Y|I) = P(X|Y, I)P(Y|I). \tag{5.4}$$

All of the above probabilities are conditioned on some background information I to remind ourselves that in real life we always have some kind of previous knowledge upon which we base our beliefs. Jaynes [1] in particular makes a big fuss in his opening remarks about how this previous knowledge is often ignored by researchers and not put to good use to better their analysis. The I will often be omitted later but it will always be implied. One might ask why Cox's argument should lead to the usual rules of probability, and the answer is that he derived the results by considering boolean logic and imposing the consistency requirement that any method of analysis given the same information should always lead to the same conclusion. For a more rigorous derivation see appendix B of Sivia [2] and chapter 2 of Jaynes [1].

5.2.3 Bayes' Theorem

A simple consequence of the product rule 5.4 is Bayes' theorem. Notice that the joint probability $P(X, Y|I)$ has to be the same as $P(Y, X|I)$, so we must have that

$$P(X|Y, I)P(Y|I) = P(Y|X, I)P(X|I). \tag{5.5}$$

A simple rearrangement yields

$$P(X|Y,I) = \frac{P(Y|X,I)P(X|I)}{P(Y|I)}. \tag{5.6}$$

This theorem is the crux of logical inference and Bayesian analysis (hence the name) because it reverses the dependence of the variables. In particular, given some experimental data and a theory or hypothesis that we want to test, we would estimate the probability that our hypothesis is true as follows:

$$P(hypothesis|\ data, I) = \frac{P(data|\ hypothesis, I)P(hypothesis|\ I)}{P(data|\ I)}. \tag{5.7}$$

Or, more generally,

$$Posterior = \frac{Prior\ \times\ Likelihood\ Function}{Evidence}. \tag{5.8}$$

The posterior probability is the probability that our hypothesis is true, given the data and prior information. In other words, it is our inference from the observations as to how much we believe a certain proposition. The situation is not dissimilar to that of a jury trying to decide whether the defendant is guilty or innocent after being presented with evidence from the prosecution and a counter argument by the defence. If the defendant did indeed stab the victim, then there is a good chance that the victim's blood would splatter onto defendant's clothes. So if we do indeed find blood on the defendant's shirt, we would be inclined to believe that he or she did commit the crime, but we would not be 100% sure as there might be other explanations. Similarly in Bayes' theorem, the likelihood function computes the probability of obtaining the observed data if the hypothesis were true, and, along with the estimates for the other two probabilities, gives us an estimate of whether the hypothesis is then likely to be true based on the data.

5.2.4 Marginalization

Another consequence of the sum and product rules is the idea of marginalization. Consider two propositions X and Y. Adding the probabilities $P(Y,X|I)$ and $P(\overline{Y},X|I)$ and using the product rule we get

$$\begin{aligned} P(Y,X|I) + P(\overline{Y},X|I) &= P(Y|X,I)P(X|I) + P(\overline{Y}|X,I)P(X|I), \\ &= [P(Y|X,I) + P(\overline{Y}|X,I)]P(X|I), \\ &= P(X|I). \end{aligned}$$

Since the order of the joint probabilities doesn't matter, we also have that

$$P(X|I) = P(X,Y|I) + P(X,\bar{Y}|I). \tag{5.9}$$

This expansion is rather interesting as it eliminates the Y dependence in the two joint probabilities and leaves behind only the probability of X alone. In general, suppose we had a set of mutually exclusive and exhaustive propositions $\{Y_k\}$, $k = 1, 2, ..., M$ so that

$$\sum_{k=1}^{M} P(Y_k|X,I) = 1, \tag{5.10}$$

it will be true that

$$P(X|I) = \sum_{k=1}^{M} P(X,Y_k|I). \tag{5.11}$$

So in the continuum limit, we have

$$P(X|I) = \int_{-\infty}^{\infty} P(X,Y|I)dY, \tag{5.12}$$

where $P(X,Y|I)$ is now interpreted as the probability density rather than the probability. Marginalization is actually a very powerful tool because it allows us to eliminate *nuisance parameters* - parameters that we do not know or do not care about. Imagine that you wanted to estimate the average height of a first year student. It is reasonable to assume that heights are normally distributed but we do not know anything about the variance of the heights, so we would have to estimate it as well. Since we do not care about the spread of the heights, we can use marginalization to eliminate variance from our model, and estimate just the mean. Sivia [2] explains how to employ marginalization to remove background noise from a signal in chapter 3 of his book.

Chapter 6

Estimating the HP5

Armed with the basic tools of probability theory, we can now actually set out to estimate the HP5 in a very simple and straightforward manner. Recall that HP5 is defined as percentage of readings under 5 mmHg. So all we care about is whether the measurement is above or below 5 mmHg and the actual numerical value of the reading doesn't matter. With each piece of data reduced to only two possible outcomes, what we have here is simply a Bernoulli trial, with success meaning the reading is less than 5 mmHg and failure meaning it isn't, and HP5 is really just intuitively the probability that any given reading is below 5 mmHg. (A tacit assumption we have made here is that each reading is independent, which seems plausible, though may not be the case given how the needle probe advances through the tissue.) To analyze the very general situation of finding the probability of success in a Bernoulli trial, let us consider the concrete example of flipping a coin.

6.1 Motivation: Coin Flipping - How do you know if a coin is fair?

Suppose that we have picked up some coin of unknown origin. Assuming that we, or whoever flips the coin, do not have any special way of flipping a coin to generate a certain outcome, what is the probability of getting heads on any flip? If the coin were fair then we would know that the probability for heads is 1/2 and similarly the probability of getting tails is 1/2, but in practise one would not know that a priori, and the only thing that can be done is to flip the coin many times, and try to guess from the data whether the coin is fair or not. That is, we want to estimate the probability of getting a heads (and hence the probability of getting a tails) on

any given flip.

To that end, let's assume that we have flipped the coin several times and collected some data, which we denote as $\{D\}$. $\{D\}$ can be many flips, say several hundred, or it can be as few as one, all that will change is the level of confidence we have in the outcome of our analysis. Now define H to be our bias factor (which is just a fancy name for the probability of obtaining heads), with $0 \leq H \leq 1$, and we wish to estimate H, or more precisely H given our data $\{D\}$, denoted by $P(H|\{D\}, I)$. As Bayesians, we will of course appeal to Bayes' theorem:

$$P(H|\{D\}, I) = \frac{P(\{D\}|H, I)P(H|I)}{P(\{D\}|I)}. \tag{6.1}$$

It must be stressed here that the probability that we are trying to obtain, $P(H|\{D\}, I)$, is the probability distribution *for* H that encapsulates *our degree of belief* for any value, or range of values, of H given our data $\{D\}$. It in no way means that we have assumed H to be a random variable. We have, in fact, assumed quite the contrary; we have assumed that H is fixed. The thing is, we do not know what the value is so the best we can do is say that we believe it to be a certain value to a certain degree of confidence, expressed in the form of a probability distribution.

6.1.1 Prior Probability

Assuming that we have no prior information that leads us to favor any value, or range of values, of H over any other, the prior that correctly expresses our complete ignorance will be the uniform prior:

$$P(H|I) = \begin{cases} 1 & \text{if} \quad 0 \leq H \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

In other words, we are positing that all values of H between 0 and 1 are equally likely because we have no reason to believe otherwise. I think that the assignment of the uniform prior, given that we have no prior information about H, is plausible and reasonable, and a more formal mathematical justification is given in Appendix E (known as the *principle of indifference*). If we had some other information about what the value of H was, then we would have to choose a different prior that reflects this extra knowledge. How we choose that prior will depend on the type and specificity of the information that we are given, and there is extensive debate and much controversy over what is the correct way to do this. The most common

way of assigning probabilities, the principle of maximum entropy, is the focus of the Appendix E.

6.1.2 Likelihood Function

Since the only possible outcomes are heads or tails, and assuming that the flips are independent of each other, it seems intuitively obvious that the likelihood function, which is the probability of getting the observed data given a certain value of H, should be the binomial distribution:

$$P(\{D\}|H,I) = \binom{n}{r} H^r (1-H)^{n-r}, \tag{6.3}$$

where n is the total number of tosses or trials, and r is the number of heads or successes recorded. A formal derivation of this pdf is given in Appendix D; we will simply take it for granted here that the likelihood function must be binomial.

6.1.3 Evidence

Since the evidence term, $P(\{D\}|I)$, does not depend on the bias weighting H, it is constant with respect to H, and will simply act as the normalization constant

$$P(\{D\}|I) = c. \tag{6.4}$$

The evidence term is, in general, constant in situations such as the coin flip where we are trying to estimate an unknown parameter.

6.1.4 Posterior

Putting everything together and applying Bayes' theorem (5.6), we get for the posterior

$$P(H|\{D\},I) = \begin{cases} C \cdot H^r (1-H)^{n-r} & \text{if} \quad 0 \leq H \leq 1, \\ 0 \quad \text{otherwise,} \end{cases} \tag{6.5}$$

where C is a constant chosen to normalize the probability. Normally, C will be determined numerically, but it turns out that in this case we can determine C analytically so I will do so as follows. We will need this result later in Appendix G when looking at Laplace's rule of succession.

To normalize the probability distribution, we need C to satisfy the following condition:

$$\int_0^1 C \cdot H^r(1-H)^{n-r}dH = 1. \tag{6.6}$$

The integral on the left hand side is a standard integral known as the beta function and has the following nice form:

$$\int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{6.7}$$

Combining (6.6) and (6.7) we get

$$C = \frac{(n+1)!}{r!(n-r)!}. \tag{6.8}$$

So the posterior distribution for the bias parameter H is in fact

$$P(H|\{D\},I) = \begin{cases} \frac{(n+1)!}{r!(n-r)!}H^r(1-H)^{n-r} & \text{if} \quad 0 \leq H \leq 1, \\ 0 \quad \text{otherwise,} \end{cases} \tag{6.9}$$

where n and r would be plugged in from our data. This simple equation is our best guess as to whether the coin is fair, and even though it looks exactly like the likelihood function, the variable here is H, whereas before we were assuming H was fixed, and n and r were the variables. Equation 6.9 is more precisely the probability density function (pdf) for H given our data $\{D\}$, and it is a function that succinctly encapsulates how much we believe each value of H between zero and one to be the true value of H. The higher the value of $P(H_0|\{D\},I)$ for some fixed H_0, the more we believe H_0 to be the actual value of H, and the value of H_0 that maximizes the pdf is our best estimate.

Example. *To see what this analysis yields, suppose we had flipped a coin 50 times and it had come up heads 37 times. Such a result would lead us to believe that the coin is biased towards heads since heads came up almost three times more than tails. Using our previous analysis we would get that*

$$P(H|\{D\},I) = \begin{cases} \frac{(51)!}{37!(13)!}H^{37}(1-H)^{13}, & \textit{if} \quad 0 \leq H \leq 1, \\ 0, \quad \textit{otherwise.} \end{cases} \tag{6.10}$$

This equation isn't too illuminating on its own, but looking at the plot (figure 6.1 of the probability distribution for H confirms our suspicions that the coin may be biased because the probability that the coin is fair, $H = 0.5$, is almost zero, and that the more probable values of H are all larger than 0.5, with the maximum at about $H = 0.75$.

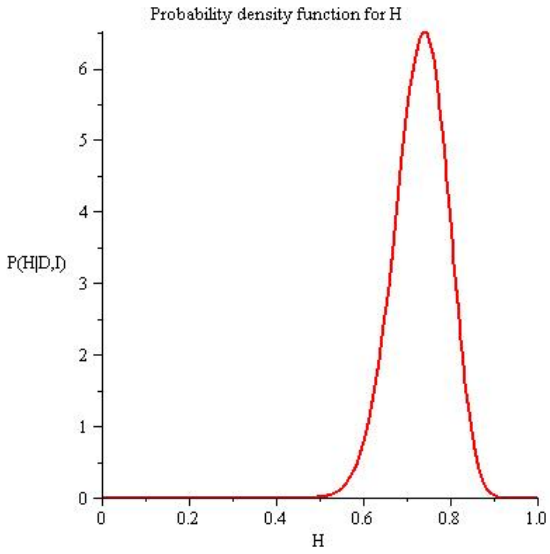


Figure 6.1: A plot of the pdf in example 6.1.4. Based on this plot (i.e on the data), we would not believe the coin to be fair.

As explained earlier, the preceding analysis is equally valid for estimating the HP5 of a tumour, so I will be using the result, equation 6.10, without modification to estimate the HP5 of a patient based on the oxygen probe measurements for that patient. Before presenting the results, there are some details regarding the pdf that we need to consider.

6.2 Sequential data analysis

In practice, data is likely to be accrued over time, so one might ask: should we wait until we have all the data, or can we analyze what we have already, and update our analysis accordingly as more data comes in? In many cases, the distinction is meaningless because all the data is collected within a short time span, but for some situations, such as survival data, we could be waiting a long time for all the data to be collected, so the second option is preferable.

Consider the case where we only have two pieces of data D_1 and D_2. Recalling that the evidence term is a constant in parameter analysis, we have

$$P(H|D_1, D_2, I) \propto P(D_1, D_2|H, I)P(H|I), \tag{6.11}$$

where $P(D_1, D_2|H, I)$ would be the joint probability of observing both D_1 and D_2,

and corresponds to the situation where we plug in all the data at once. However, we do not necessarily have to expand $P(H|D_1, D_2, I)$ in the manner above; we could also subsume D_1 into the background information

$$P(H|D_1, D_2, I) \propto P(D_2|H, D_1, I)P(H|D_1, I). \tag{6.12}$$

If the data were independent (which we are assuming to be so), then the first term on the right hand side can be written as

$$P(D_2|H, D_1, I) = P(D_2|H, I), \tag{6.13}$$

since the observation of D_1 has no bearing on whether D_2 is observed on not. Furthermore, the term on the far right is the posterior probability for H having observed D_1, so it can be rewritten as

$$P(H|D_1, I) \propto P(D_1|H, I)P(H|I). \tag{6.14}$$

Thus we have

$$P(H|D_1, D_2, I) \propto P(D_2|H, I)\left[P(D_1|H, I)P(H|I)\right], \tag{6.15}$$

which in words means that, if D_1 and D_2 are independent, then the posterior probability for H given both those pieces of data is simply the product of the posterior for H given D_1 and the likelihood function for observing D_2. That is, we can treat the two pieces of data as though we had only observed D_1, calculated the posterior for H to see what kind of estimate that gives us, then observed D_2, and updated our posterior by multiplying the the likelihood function for D_2 with the posterior for D_1. We can think of the posterior for H given D_1 as the prior probability for D_2. It should be clear that if we had more data, we would simply repeat the previous argument inductively, breaking down the joint probability by taking out one piece of data at a time. Figure 6.2 illustrates how our posterior changes as we observe more data.

6.3 Effect of Different Priors

We have assumed the situation of total ignorance, but there are many situations where we will have access to information about the quantity of interest, and that will lead us to assign a more informative prior than the flat prior. To see the effects of different priors, suppose we had some extra information about HP5. Say

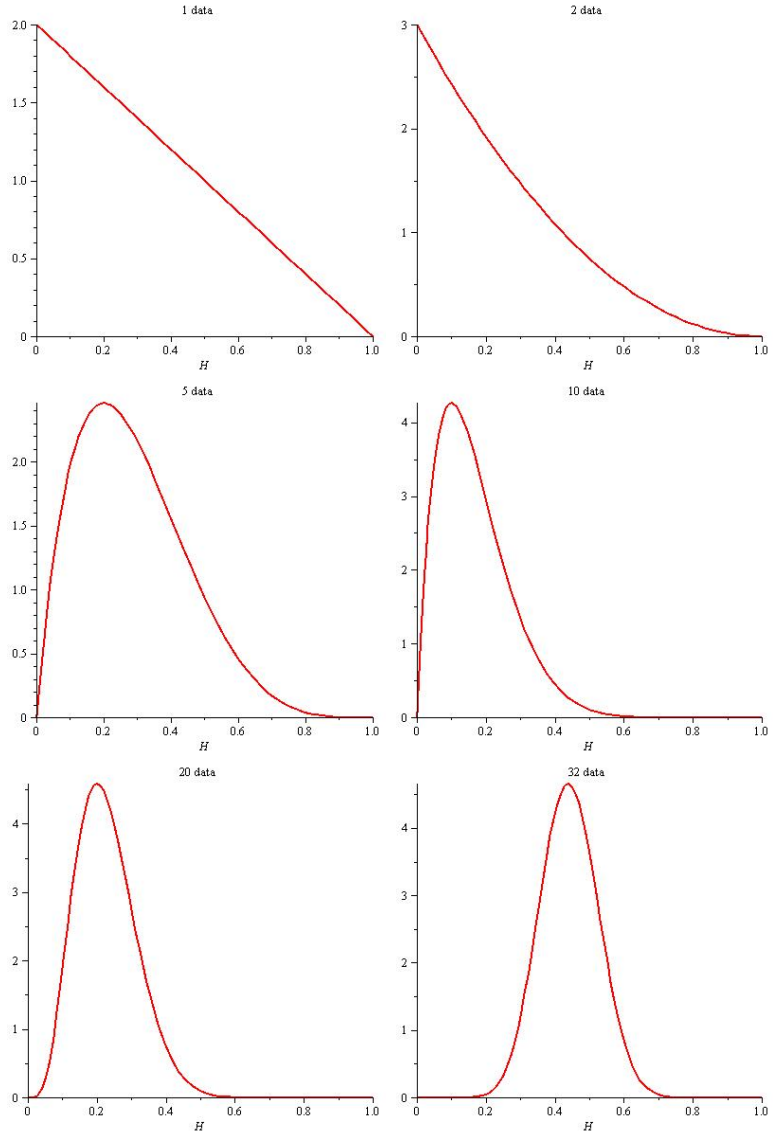


Figure 6.2: Estimating the HP5 for patient P19 by building up the posterior sequentially. As we accrue more data, we simply multiply the new likelihood function with the posterior we had previously to generate an updated posterior probability distribution. Notice how as we add more data the posterior probability becomes narrower, indicating that our error is decreasing as we hone in on a better estimate of H_0.

we knew that for the average cervical tumour we expect HP5=0.8. It will turns out that, using the principle of maximum entropy outlined in Appendix E, having information about expectation, or the mean, alone will lead to an exponential prior

$$P(H|\mu) = \frac{1}{\mu} e^{-\frac{H}{\mu}}, \tag{6.16}$$

where μ is the known average value, and knowing the mean and the variance will lead to a Gaussian prior

$$P(H|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(H-\mu)^2}{2\sigma^2}}, \tag{6.17}$$

where σ^2 is the known variance. The situation is illustrated graphically in figure 6.3, and in short, what happens is that the prior will play a large role if there is little data, but will become largely irrelevant as we consider more and more data. This makes sense logically as we would be less likely to reject our previous knowledge of the situation if there were little evidence, and more likely to disbelieve it as we observe more evidence to the contrary. More specifically, comparing the Gaussian prior to the other two priors, we see that the more specific our information, the narrower the prior probability distribution is, and the harder it is to dissuade ourselves that our prior information is incorrect.

6.4 Summarizing the Results

Having the actual probability distribution that expresses our degree of belief for the parameter value is good from a mathematical stand point, and there are methods for summarizing the key features of the distribution in cases where the posterior is reasonably well behaved. I will discuss two sets of methods: the best estimate and the error bar, a method that works best with symmetric pdfs; and the expectation value and confidence interval, a similar technique that is more appropriate for asymmetric distributions.

6.4.1 Best Estimate/ Maximum Likelihood

As discussed earlier, the value of H_0 that we believe the most is the value that gives us the highest probability in $P(H_0|\{D\}, I)$. More mathematically, suppose we were interested in some quantity X, and we had the posterior pdf $P(X|\{D\}, I)$. Then the *best estimate* of X, denoted by X_0, is determined by

$$\left.\frac{dP}{dX}\right|_{X_0} = 0. \tag{6.18}$$

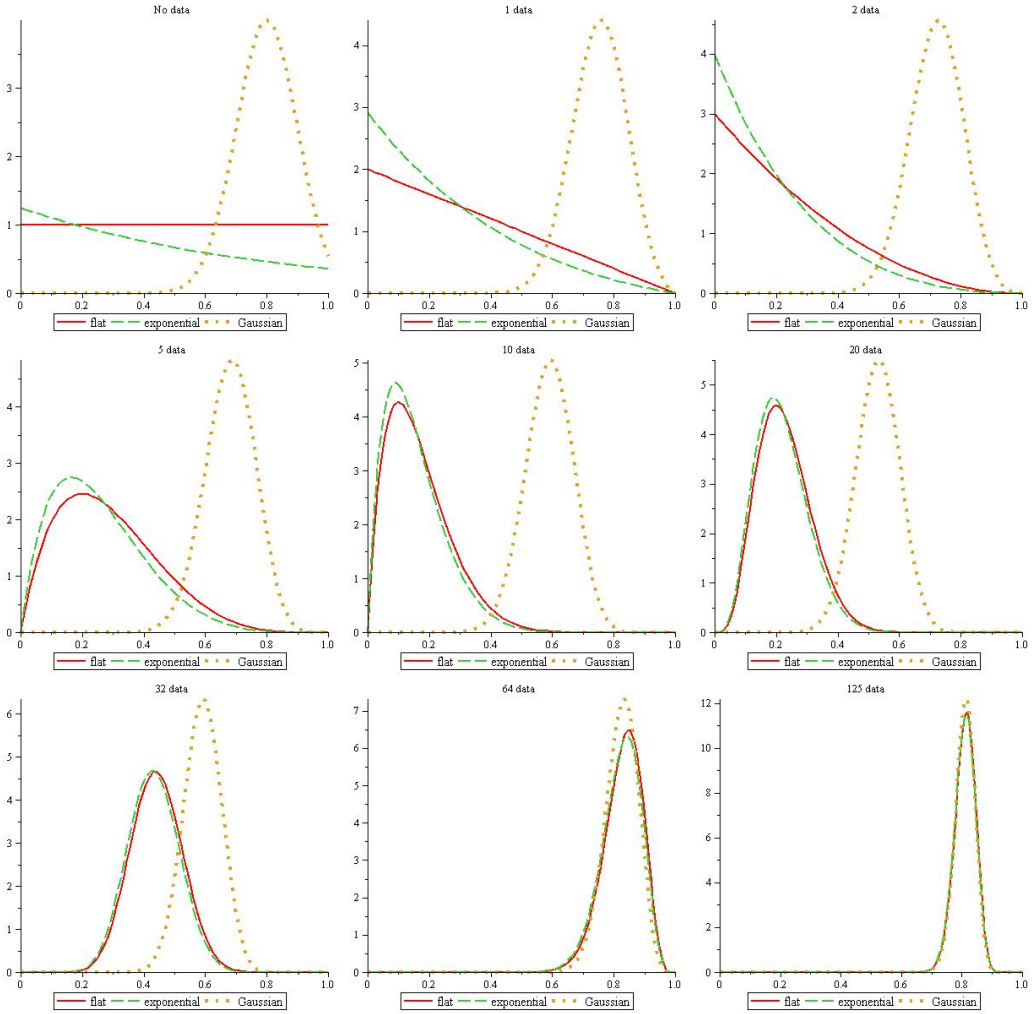


Figure 6.3: The effects of different priors on the HP5 estimate for patient P19. From the plots, we see that, since the flat prior and the exponential prior are fairly similar, their posteriors become almost identical after just 20 pieces of data. However, because the Gaussian prior is tightly peaked about a value that is in disagreement with our experimental data, it takes much more evidence to "persuade" the posterior to believe the data over the prior information, and we can see that the pdf for the Gaussian prior does in fact slowly converge to the other two pdfs as we increase the amount of data.

Of course, we could try to verify analytically that this value of X_0 actually yields a maximum, but in practice we will simply look at a plot of the pdf. If we were considering discrete values for X instead, the best estimate is still the value that gives the largest posterior probability, but the derivative notation would not make sense. This method of determining the best estimate is often called the maximum likelihood method because in the case of parameter estimation, if the prior is the flat indifferent prior, maximizing the posterior is the same as maximizing the likelihood function.

6.4.2 Error Bar/ Variance

Once we have the best estimate, we want to know how accurate that estimate is, i.e. how much we believe it to be the true value of H, and that is accomplished by considering the spread of the posterior about the best estimate. We traditionally work with the logarithm of the posterior, which varies more slowly with X than the posterior, and take the Taylor series about the best estimate X_0. Define:

$$L = ln(P(X|\{D\}, I). \tag{6.19}$$

Expanding about X_0 we get

$$L(X) = L(X_0) + \frac{1}{2}\left.\frac{d^2L}{dX^2}\right|_{X_0}(X - X_0)^2 + ..., \tag{6.20}$$

where the order-one term has disappeared since we are expanding around the maximum, and $L(X_0)$ is just a constant. So if we are close to X_0, we can drop the higher order terms and get

$$P(X|\{D\}, I) \simeq Aexp\left[\frac{1}{2}\left.\frac{d^2L}{dX^2}\right|_{X_0}(X - X_0)^2\right], \tag{6.21}$$

where A is the normalization constant. Upon closer inspection, this expression is actually just the normal distribution:

$$P(H|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(H-\mu)^2}{2\sigma^2}}. \tag{6.22}$$

Comparing the two we see that the variance must be related to L through the following relationship:

$$\sigma = \frac{1}{\sqrt{\frac{d^2L}{dX^2}\big|_{X_0}}}. \tag{6.23}$$

Note that although we denote this quantity by the symbol used for variance, and it is also sometimes referred to as such, it is not the same as the sample variance, which is normally estimated by squaring equation 4.4:

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{6.24}$$

The σ in our case is normally called the *error bar* and is a measure of the range of values that we expect the true value of X to reside in. In other words, it describes the accuracy of our best estimate.

We normally summarize our posterior pdf by the statement:

$$X = X_0 \pm \sigma, \tag{6.25}$$

because, from equation 6.21, we expect that

$$P(X_0 - \sigma \leq X < X_0 + \sigma | \{D\}, I) = \int_{X_0-\sigma}^{X_0+\sigma} P(X|\{D\}, I)dX \simeq 0.67. \tag{6.26}$$

That is, we are 67% sure that the true value of X is in the range $[X_0 - \sigma, X_0 + \sigma]$, and usually that is a good enough guess for X. What we have done here is to approximate our posterior with a Gaussian for values near the maximum; for posteriors where the central limit theorem holds, this will be a good approximation and Sivia [2] provides an interesting counter example where that isn't the case in chapter 2 of his book.

Back to the coin

For the case we are considering, where we use a flat prior and the likelihood function is the binomial distribution, it can be shown that

$$H_0 = \frac{R}{N}, \qquad \sigma = \sqrt{\frac{R}{N^2}\left(1 - \frac{R}{N}\right)}. \tag{6.27}$$

The thing to note here is that the best estimate, H_0, is the simple ratio of number of measurements under the threshold divided by the total number of measurements, and is the value that we would have gotten had we just followed Bernoulli's method. However, by applying Bayesian inference, we not only have that estimate of HP5, but a measure of how much we believe it to be true as well.

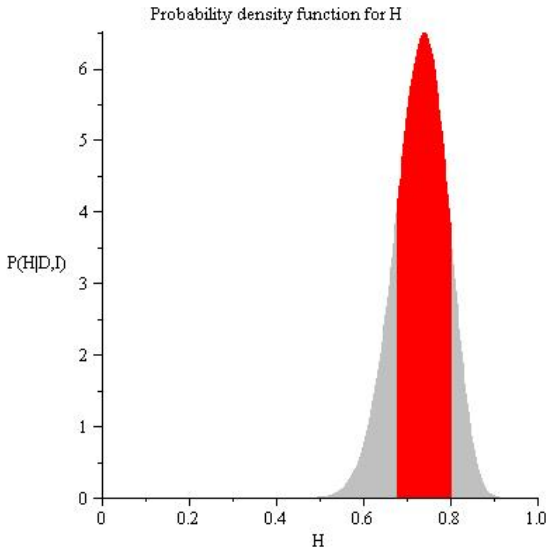


Figure 6.4: The darker region shaded in red is the width of $H_0 \pm \sigma$ for the previous coin example, which is approximately 67% of the area under the curve, and tells us that we are 67% sure that H is within that range. In terms of actual values, we get $H_0 = 0.74 \pm 0.062$. Note that although 0.74 is the same as the sample mean, the sample variance is 0.2 and is a measure of the spread of the data, whereas the error bar is a measure of our confidence in the best estimate H_0.

6.4.3 Mean/ Expectation Value

If the posterior is skewed to one side of the peak, specifying a symmetric interval about the peak will make little sense; in this case, we often replace the maximum with the expectation or mean value

$$< X >= \int XP(X|\{D\}, I)dX. \tag{6.28}$$

The expectation is basically just a weighted average that takes into account the skewness, and reduces to the maximum when the posterior is symmetric.

6.4.4 Confidence Interval

To replace the error bar, we find the confidence (or credibility) interval, which is the smallest interval (X_1, X_2) such that

$$P(X_1 \leq X < X_2|\{D\}, I) = \int_{X_1}^{X_2} P(X|\{D\}, I)dX \simeq 0.95. \tag{6.29}$$

Here 95% is simply the standard choice, and is equivalent to $X_0 \pm 2\sigma$ in the symmetric case.

6.4.5 Multimodal Posterior Distributions

In the previous discussion, we have implicitly assumed the posterior to only have one well defined peak, but there will be situations where there is more than one peak. For multimodal posteriors the best estimate is still meaningful when one peak of the pdf is much higher than the others, as it still represents the value of the parameter that we believe the most, but the other methods for summarizing the results are inappropriate since they all implicitly assume one nice behaved peak.

For example, consider a distribution with two peaks of equal size, spaced symmetrically about the middle of the interval (see figure 6.5). Specifying the best estimate and error bar is dubious because we would have to say our best estimate is the peak on the left $\pm\sigma_{left}$ OR the peak on the right $\pm\sigma_{right}$, which is confusing and not too helpful. The expectation is worse as it would give us the value in the middle, which from the plot of the pdf we know has a probability of close to zero, and similarly we have no hope of specifying the confidence interval in a sensible manner. In complex situations, the best we can do is plot the posterior and extract information as necessary.

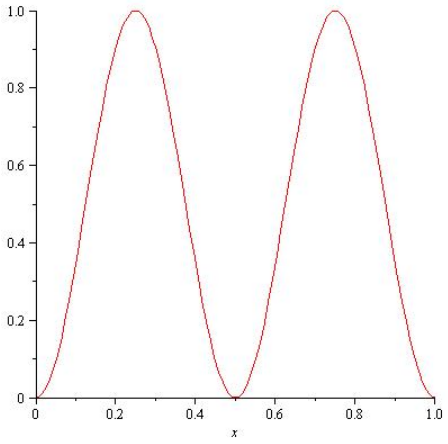


Figure 6.5: An example where our methods for summarizing the posterior distribution with just a couple of numbers make no sense.

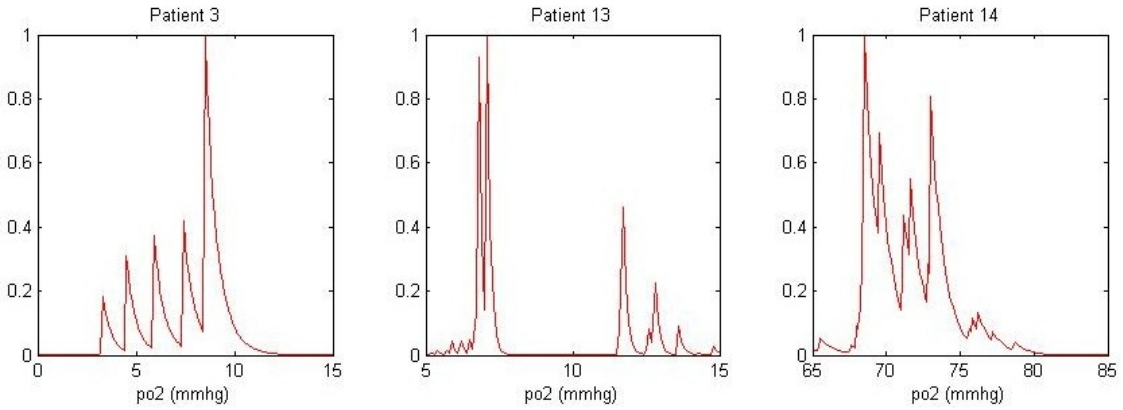


Figure 6.6: In the next chapter we will estimate the median PO2 using maximum entropy methods, and the analysis will yield posteriors that are multimodal (shown above). For patient 3, we can still say that the most probable value is the one given by the maximum of the distribution, as there is clearly one higher spike than the others, but the maximum is not so clear for the other two patients because there are several values that are almost equally probable. The best thing to do in cases such as these is to simply display the pdfs as shown here.

6.5 Results

We will use the equation for the posterior distribution of H, equation 6.9, to estimate the HP5 of each patient based on the Eppendorf measurements. To do so, we will first turn the Eppendorf data into binary data (either 1 if PO2<5mmHg and 0 otherwise), add them up and then plug in the appropriate values into equation 6.9. The binomial distribution is symmetric about its peak, so we will specify the best estimate and error bar, given by equation 6.27. Figure 6.8 is a table that shows the results for some of the patients, and figure 6.9 displays the probability distributions graphically.

6.5.1 Possible Advantages

In the original study [11] the value of HP5 is just a number, a ratio of two other numbers that offers one key piece of information, an estimate of the hypoxia, but not much else. It was then used to classify patients into two groups of varying survival probabilities. By applying Bayesian statistics, we have preserved that original piece of information in our best estimate H_0 and generated new information via the

Study Number	Track Number	Position Along Track	PO2	<5mmHg?
1	1	1	-1.3	1
1	1	2	15.2	0
1	1	3	9.7	0
1	1	4	30.6	0
1	1	5	37.7	0
1	1	6	19.1	0
1	1	7	8.5	0
1	1	8	9.4	0

Figure 6.7: To calculate the HP5, we first ask ourselves whether each measurement is above or below 5 mmHg. Then, we tally the results and plug the appropriate values into equation 6.9.

probability distributions derived for each patient. At first glance this difference seems purely academic and of no practical importance as using the best estimate alone would not change any of the survival analysis at all, but I believe there are advantages to the Bayesian formulation.

The Accuracy of our Best Estimate

One key feature of the Bayesian analysis is that we were able to derive a quantity, the error bar, that measures the uncertainty of our best estimate. The original formulation of HP5 just gives us a number, we have no idea how much we can trust that estimate, and we are not able to tell whether the estimate was a result of only a few measurements, or many measurements. The error bar is a function of the number of measurements and reflects how well the data agrees with our best estimate.

One possible use of the error bar is to use it to analyze how the error changes with the number of tracks or measurements. We could possibly use it to quantify how many readings on average we would need to reduce the error to a certain amount, giving us an idea of how many measurements we would need to be reasonably sure of how hypoxic the tumour is.

	HP5 estimate			
Patient	One Track	Two Tracks	Three Tracks	Four Tracks
109	0.947 ± 0.052	0.745 ± 0.061	0.711 ± 0.050	0.513 ± 0.047
110	0.931 ± 0.047	0.947 ± 0.030	0.663 ± 0.050	0.587 ± 0.045
111	0.824 ± 0.092	0.854 ± 0.051	0.863 ± 0.040	0.860 ± 0.035
P01	0.321 ± 0.062			
P02	0.071 ± 0.068	0.022 ± 0.022		
P04	0.469 ± 0.088	0.314 ± 0.065	0.414 ± 0.059	0.523 ± 0.054
P07	0.250 ± 0.097	0.122 ± 0.051	0.194 ± 0.050	
P08	0.375 ± 0.086	0.565 ± 0.063	0.696 ± 0.048	0.597 ± 0.044
P11	0.810 ± 0.086	0.643 ± 0.074	0.450 ± 0.064	0.333 ± 0.052
P12	0.619 ± 0.106	0.465 ± 0.076	0.639 ± 0.057	0.663 ± 0.050
P16	0.583 ± 0.101	0.607 ± 0.065	0.523 ± 0.053	0.550 ± 0.045
P19	0.438 ± 0.088	0.656 ± 0.059	0.771 ± 0.043	0.816 ± 0.035

Figure 6.8: Table of best estimates and error bars for some patients as we increase the number of tracks. Not all patients have four tracks of data so some of the columns are blank. The original HP5 estimate used by Fyles et al would correspond to our best estimate when all the data is used (i.e. the right most estimate in this table).

Stratification into groups for testing survival

Another possible use for the error bar is in survival analysis. Recall that to analyze survival, Fyles et al divided the patients into two groups based on their HP5. Now that we have a sense of how accurate our estimate of the HP5 is, we can be more flexible in defining the groups. For example, in table 6.8, patient P04 has a HP5 estimate of 0.523 ± 0.054. If we did not have the error bar we would simply classify this patient as hypoxic (HP5>.5) but given that our analysis reveals that the true value of HP5 might be as low as $0.523 - 0.054 = 0.469$, we might reasonably classify the patient as oxic as well. As shown in Appendix F using filters, being able to reclassify these patients that are on the border between the two groups can yield different results in the survival analysis, so having this extra flexibility might be beneficial.

Incorporation of Prior Knowledge

Another possible advantage of the Bayesian approach is the ease with which we can incorporate any previous knowledge we might have. We saw earlier in the

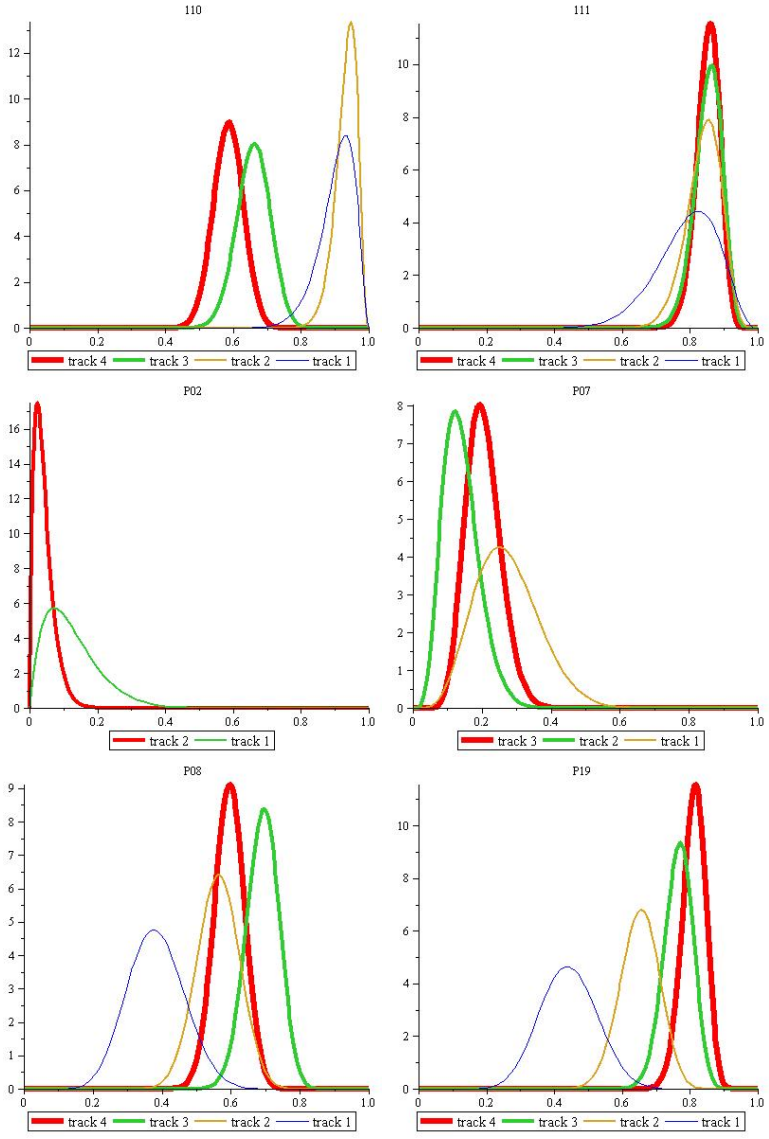


Figure 6.9: Plots of some of the probability distributions as we increase the amount of data (number of tracks). Again notice how the distributions tend to become narrower as we increase the data, indicating that we have increased confidence in our best estimate.

comparison of priors how extra information can used to assign more informative priors, and that information might be available from other researchers or from the manufacturer of the Eppendorf probe. For instance, we might have information from past studies that patients of a certain age, height, tumour size etc would be likely to have a certain HP5. Or, Eppendorf might tell us that their probe has a certain accuracy in human tissue, and we could incorporate that into our model. At the moment, we have assumed that the measurements are accurate even though there are obviously impossible negative values.

Probability Distributions as Opposed to One Number

Having access to the probability distribution gives us added flexibility as we can manipulate it to calculate other quantities of interest. One quantity that we might want to calculate is the probability that a patient has a HP5 of less than some threshold, and use that as our estimate of the hypoxia instead of the best estimate H_0. For example, for patient P08, we have $H_0 = 0.597$, and the probability that the value of HP5 is less than 0.5 is 0.157.

Chapter 7

Estimating the Median of the PO2 Distribution

The median PO2 is another quantity of interest because it is a commonly used measure of hypoxia and also because Fyles et al [11] used the sample median of 5 mmHg in their data for their threshold in hypoxic proportion. Using the method of maximum entropy outlined in the Appendix E, we can estimate the median of the unknown underlying distribution that our data is a sample of, and see how that compares to the sample median.

7.1 Theory

When we perform an experiment, say rolling a die or conducting polls etc, what we are doing is collecting a sample from some underlying probability distribution, or population, of possible outcomes. This distribution is generally unknown, and what orthodox statisticians usually do is try to use the samples that have been collected to estimate the distribution. Estimating the distribution can be very difficult, especially if we do not have any idea of its functional form, so what we will try to do here is to estimate the value of the median (or more generally, the value of some quantile) of the distribution.

It must be stressed that the median of the distribution is (in general) not the same as the median of the samples. To calculate the sample median, all we would do is line them up in order and pick the one in the middle. However, this value will be vary depending on the samples that have been collected and is not necessarily

a good estimate of the median of the population.

Mathematically, we wish to estimate the position of a certain quantile, q, of the pdf $P(x)$, $x \in [a, b]$, of the quantity that we are interested in. That is, we want to find x such that

$$\int_a^x P(x')dx' = q, \tag{7.1}$$

where $a \leq x \leq b$, and $0 \leq q \leq 1$. Now suppose we had N samples, $\{x_i\}$, taken from the probability distribution $P(x)$. Let M_x be the hypothesis that the median is at x:

$$M_x : \quad \int_a^x P(x')dx' = 1/2. \tag{7.2}$$

Then what we want to calculate is the posterior probability for M_x given our data. Using Bayes' theorem we have

$$P(M_x|\{x_i\}) \propto P(M_x)P(\{x_i\}|M_x). \tag{7.3}$$

Since we have no reason to believe any value of x over another for the median, all we know a priori is that the median must lie between a and b. So for the prior, $P(M_x)$, we will simply assume a uniform or constant probability in the interval $[a, b]$ and zero everywhere else.

For the likelihood function, $P(\{x_i\}|M_x)$, we know nothing about its functional form, other than the constraint that its median must be at x, so we must use some kind of non-parametric analysis. Fortunately, since the hypotheses M_x are testable, we can make use of the principle of maximum entropy to assign the likelihood function in a way that satisfies our constraints but still reflects our ignorance of everything else. We will assume further that the samples are exchangeable so that M_x is testable information about the joint pdf $P(\{x_i\})$. Following Gull and Fielden [8], we wish to maximize

$$S = -\int P(\{x\})ln\left(\frac{P(\{x\})}{m(\{x\})}\right)d^N x, \tag{7.4}$$

subject to the following constraints:

$$\int_D P(\{x\})d^N x = 1, \quad \text{where } D \text{ is the domain } [a, b]^N, \tag{7.5}$$

$$\int_{D_x} P(\{x\})d^N x = \frac{1}{2}, \quad \text{where } D_x \text{ is the domain } [a, x]^N. \tag{7.6}$$

In the case of a single sample, with a uniform measure $m(x)$, the analysis simplifies, and we get [8]:

$$P(x'|M_x) = \begin{cases} \frac{1}{2(x-a)}, & \text{if } x' < x, \\ \frac{1}{2(b-x)}, & \text{if } x' > x. \end{cases} \tag{7.7}$$

For multiple, exchangeable, samples, with uniform measure $m(\{x\})$, the variables are independent, so we simply have

$$P(\{x_i\}|M_x) = \left[\frac{1}{2(x-a)}\right]^{N_<} \left[\frac{1}{2(b-x)}\right]^{N_>}, \tag{7.8}$$

where $N_<$ is the number of data in the set $\{x_i\}$ with $x_i < x$, and $N_>$ is the number of data with $x_i > x$. Thus, using Bayes' theorem, we must have that the posterior is given by

$$P(M_x|\{x_i\}) \propto \left[\frac{1}{2(x-a)}\right]^{N_<} \left[\frac{1}{2(b-x)}\right]^{N_>}. \tag{7.9}$$

Gull and Fielden [8] stress that this pdf is simply a maxent estimate that satisfies the constraints while assuming as little as possible about anything else, and that we are in no way claiming that it is the actual form that underlies the physical process. This is an important distinction to make, and reminds us that the posterior probability represents our own uncertainty about the value of the median x, and not that x is variable, as it is a fixed physical quantity of the system. Notice also that our assumption that $x \in [a, b]$ shows up in the final form of the posterior probability and means that our choices for a and b can potentially affect the analysis of the data.

7.2 Results

Using the Eppendorf data from the PMH group, we can apply the previous analysis to estimate the distribution for the median PO2 for each patient. From chapter 4, we know that the data we have is between -4.3mmHg and 212.8mmHg, since such extreme values of oxygen partial pressure are very unlikely in the human body, it should be safe to set the range in our prior to be $[-5, 220]$. Also in chapter 4, we saw how the Eppendorf data exhibit wildly oscillating behavior as a function of distance inside the tumour, and that would suggest that the sample median is probably not

a good estimate of the actual median PO2 in the tumour. Indeed, from our table, figure 7.1, we see that there are sometimes large discrepancies between the sample median and the maxent estimate of the median (e.g. patients 1, 2, 7, 8) but also sometimes that there is fairly close agreement (e.g patients 10, 19, 20, 24).

I am not sure as to why there are sometimes close agreements and sometimes large discrepancies between the two estimates. As we can see in figures 7.2 and 7.3 the maximum entropy estimates yield pdfs that are multimodal and that suggests our data is in disagreement with each other, with some of it suggesting one median value and others suggesting another. Perhaps this disagreeing data is because tumours are heterogeneous and have many sections of different oxygen concentrations with different median values, and the data is telling us exactly that. Thus, the best estimate of the median PO2 is probably the median PO2 of the largest region (in terms of oxygen concentrations) within the tumour, and will probably have the largest effect on radiotherapy. So whether the two estimates are in agreement or not doesn't really matter, though it does give us confidence that the Bayesian method is working when the estimates are close to each other. Perhaps in the cases where the two estimates agree, the probe is sampling a tumour that is less heterogenous and so collected data from a region that had consistent oxygenation. We might be able to check for this by looking at the plots of the data against position and seeing whether there are any large oscillations.

The new maxent estimate of the median can be used to stratify the patients into two groups as we did with the HP5. We could also use it to choose a different threshold for hypoxic proportion. I have not yet tried to use these estimates in the Kaplan-Meier and log-rank tests because the focus of this thesis was to understand the theory and produce the new estimate, and it is not yet clear how best to use the maxent estimate of the median. Also, as explained earlier, the result of the log-rank test is largely meaningless and not to be trusted.

Patient	Sample Median	Maxent Estimate
1	4.7	14.1
2	12.45	28.5
3	3.2	8.5
4	6.75	10.0
6	0.6	2.9
7	31.85	51.8
8	28.75	50.7
10	0.9	2.2
12	2.45	7.6
13	3.4	7.1
14	56.35	68.5
15	66.9	76.1
17	10.4	20.4
18	1.5	2.7
19	3.2	4.5
20	3.5	4.2
21	30.7	53.0
22	16.7	36.1
23	5.5	14.8
24	2.3	3.0
25	1.4	4.9
27	32	11.3
29	3	4.9
30	25	52.7
32	18.2	37.6
33	10.25	29.5
34	1.95	8.4

Figure 7.1: A comparison between the sample median and the maxent best estimate of the median for some patients. As the next two figures will show, the pdfs are multimodal, so there is no simple algorithm to find the maximum. To find the best estimate, I plotted the distributions and judged visually which point was the highest. The multimodal nature of the pdfs is also why I have not specified the error in the best estimate.

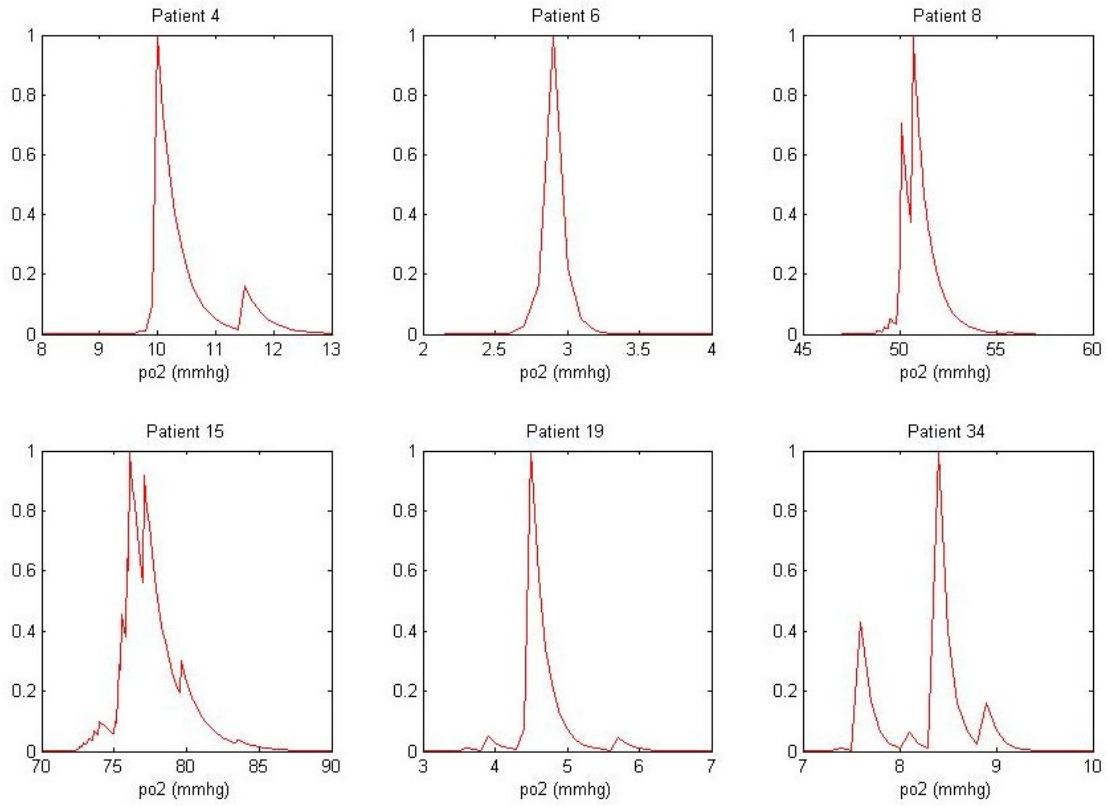


Figure 7.2: Above are estimates for the median po2 of 6 different patients that give a p.d.f. that is reasonably well behaved. By that I mean there is a clear single large peak in the distribution that would clearly indicate a possible best estimate for the value of the median, and that the probabilities for most other values are low compared to the peak. In such cases we can say fairly confidently that the median is the value at the peak with some degree of uncertainty, and might conjecture that the oxygen distribution inside the tumour is fairly homogenous as all measurements point to the same median value. (Note that the graphs are not normalized, only rescaled so that the highest point has a value of 1)

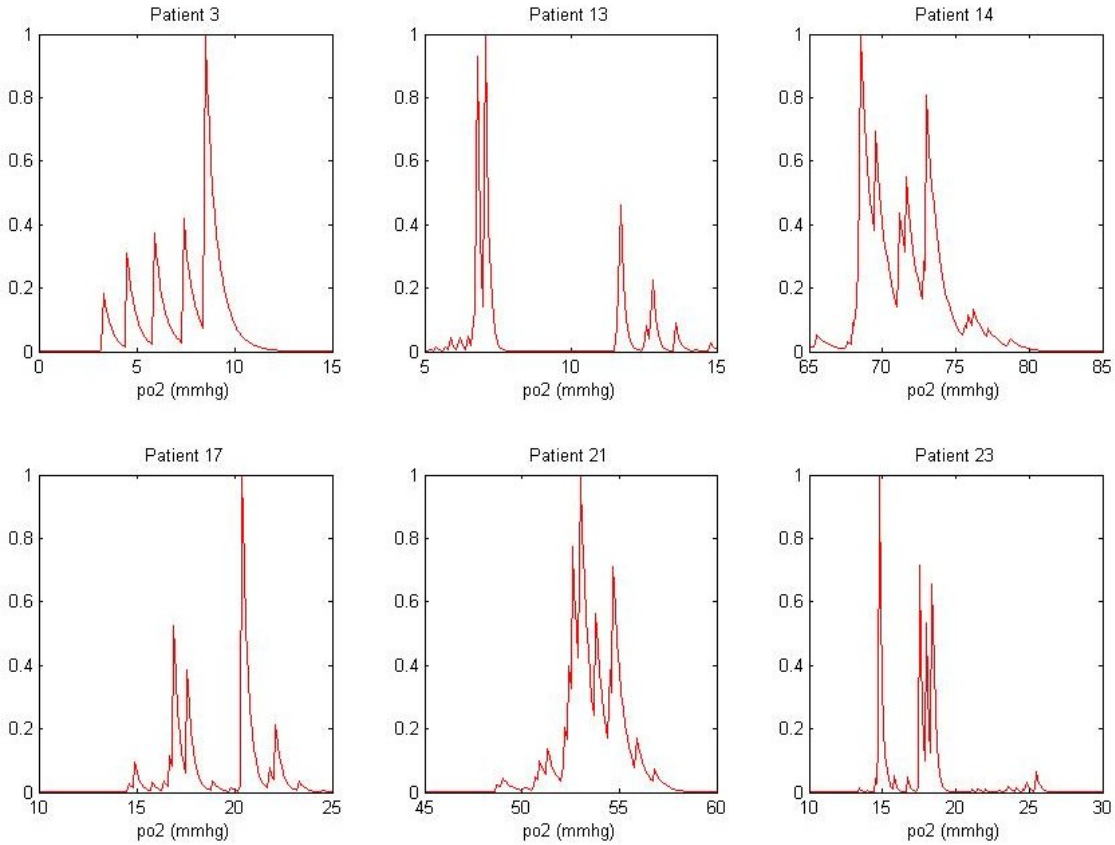


Figure 7.3: Here we have 6 patients where the p.d.f.s are not so well behaved. All the graphs shown have multiple large spikes in the distribution which means that there is no clear cut best estimate for the median. Physically, we might suspect that these tumours are fairly heterogenous but have certain regions within them that are homogenous and have a well defined median value. For instance, for patient 13, we might conjecture that the tumour has two distinct regions of oxygenation, one with a median value of about 7 mmHg and one with a median of about 12 mmHg. In such cases, what we might say is that choosing a certain threshold value and simply checking the percentage of readings under that value to make a statement about the whole tumour might not be a good idea, as there may be pockets of high and low oxygenation that such a vague analysis would not discover.

Chapter 8

Survival Analysis as an Inhomogeneous Poisson Process

Presented here in the last chapter of my thesis is a preliminary alternative to the Kaplan-Meier estimator. Although the method proposed using Poisson processes is not strictly under the Bayesian banner, we should be able to modify it in the future so that it is, since the Poisson distribution can be derived using maximum entropy (see Appendix E) under similar conditions to the ones we will outline in this chapter.

8.1 Inhomogeneous Poisson Processes

To analyze the survival data, I will assume that it is an inhomogeneous Poisson process, which is defined as follows [50]:

Definition. A point process $\{N(t), t \leq 0\}$ satisfying $N(0) = 0$ is called an *inhomogeneous Poisson process* with intensity function $\lambda(t)$ if it has the following properties:

1. $\{N(t), t \leq 0\}$ has independent increments
2. $P(N(t+h) - N(t) \leq 2) = o(h)$
3. $P(N(t+h) - N(t) = 1) = \lambda(t)h + o(h)$

The first assumption, that of independence, is a reasonable one as there is no reason to believe that the death of any individual would have an effect on the death of another patient. The second assumption means that the probability of two or more events at the same time is small, and that can be easily verified by checking the times of death and making sure that there are no patients who die at the exact same time. The third is the key mathematical assumption that leads to the poisson distribution, and is the standard assumption for counting processes, such as survival, where you count the number of events that occur in a certain amount of time (or over a certain distance/area etc).

The inhomogeneity occurs in the intensity function $\lambda(t)$, and leads to a poisson distribution for the increments that depends on t [50]:

$$p_i(s,t) = \frac{[\Lambda(t) - \Lambda(s)]^i}{i!} e^{-[\Lambda(t)-\Lambda(s)]}, \qquad i = 0, 1, 2, ..., \tag{8.1}$$

where

$$\Lambda(t) = \int_0^t \lambda(x)dx. \tag{8.2}$$

In particular, the absolute state probabilities are given by

$$\begin{aligned} p_i(t) &= p_i(0,t) = P(N(t) = i), \\ &= \frac{[\Lambda(t)]^i}{i!} e^{-\Lambda(t)}, \quad i = 0, 1, 2, ... \end{aligned} \tag{8.3}$$

Physically, we interpret the intensity function $\Lambda(t)$ as the average number of occurrences of the event (in our case death) in time t, or, equivalently, we can consider $\Lambda(t)/t$ as the average rate of occurrence per unit time. More generally, noting the way $\Lambda(t)$ appears in (8.1), we can say that

$$\frac{1}{\tau}\Lambda(\tau, t) = \frac{1}{\tau}\left[\Lambda(t+\tau) - \Lambda(t)\right] = \frac{1}{\tau}\int_t^{t+\tau} \lambda(x)dx \tag{8.4}$$

is the average rate of occurrence in the interval $[t, t+\tau]$.

From [50] we can also derive formulas for other quantities of possible interest, such as:

- $E(T_n)$, the expected time of the nth event.
- $E(Y_n)$, the expected time between the (n-1)th and nth events.
- $f(t_1, t_2, ..., t_n)$ the joint probability density of $(T_1, T_2, ..., T_n)$.

8.2 Analysis

To apply the ideas of the Poisson process we need a way to find $\lambda(x)$ for our data. Once we have determined $\lambda(x)$ we can apply the formulas of the previous section to get the probabilities of interest. For example, we might want the probability of say ten deaths within five years given our data, and all we would have to do would be to plug in the appropriate numbers.

8.2.1 Finding $\lambda(x)$

Motivated by the physical interpretation of $\Lambda(t)$, we can see that $\lambda(x)$ is sort of like the density of deaths per unit time. So to find an approximation of $\lambda(x)$ given our data, I will consider our survival data to be a particular realization of the Poisson process (see figures 8.3 and 8.4). Then, I will plot the number of deaths within certain time intervals, and fit a Gaussian through the data using least squares (see figures 8.1 and 8.2). Of course we could use other curves, but a Gaussian seems to fit the data well.

I must note here that in order to apply the analysis, I had to take out the data of 17 patients who, as noted before in chapter 4, all apparently died or suffered a relapse on the same day ($t = 0.003$). Recall that one of the assumptions behind the Poisson process was that the probability of two events occurring at the same time was essentially zero, so the fact that 17 events occurred simultaneously is clearly a violation of our assumptions. However, it seems implausible and unrealistic that so many patients all suddenly died or had a relapse at the exact same time, so I feel reasonably justified in discarding those data for this analysis.

In what follows I will abbreviate the hypoxic group, HP5>.5, as H, and the non-hypoxic group, HP5<.5, as NH. The Gaussians we get fitted to the data are (figure 8.2):

$$\lambda_H(x) = 0.14 + 13.01e^{-\frac{(x-1.15)^2}{0.82^2}}, \tag{8.5}$$

$$\lambda_{NH}(x) = 1.00 + 10.65e^{-\frac{(x-1.24)^2}{0.48^2}}. \tag{8.6}$$

We could also apply the same analysis to the data without stratification to get a different fitting that reflects the death rate of cervix cancer patients in general:

$$\lambda_C(x) = 1.14 + 22.00e^{-\frac{(x-1.22)^2}{0.71^2}}. \tag{8.7}$$

	num dead		
interval	H	NH	Total
0-1	7	2	9
1-2	11	9	20
2-3	1	1	2
3-4	0	3	3
4-5	0	1	1
5-6	0	1	1
6-7	0	0	0
7-8	1	0	1
8-9	0	1	1
9-10	0	1	1

Figure 8.1: Table for the tally of event occurrences (deaths) within uniform time intervals using the survival data.

8.2.2 Results

Now armed with the various $\lambda(x)$ we can plug them back into the formulas and take a look at the results. First, lets consider $\Lambda(t)$, the intensity function that corresponds to the expected or average number of deaths at time t (figure 8.5). From the table on the left we can see that the hypoxic group of patients has a larger average number of deaths than the non-hypoxic group, especially in the first few years, but the two eventually become almost identical as we march forward in time. This observation agrees with what we see visually in the graphs (figure 8.2), as there are more deaths in the earlier years for the hypoxic group than the non-hypoxic group, and is confirmed by the table of average rates on the right, because the rate of death in the first few years is much higher for the hypoxic group than the non-hypoxic group, and also much higher than in the later years for the hypoxic patients.

The fact that both groups eventually tend to the same value of $\Lambda(t)$ as time increases would suggest that chances of survival in the long term for both groups is about the same. Looking at the values for $\Lambda(t)$ for the cervical cancer patients as a whole and seeing a similar pattern to both of the two subgroups would also suggest that that is the case.

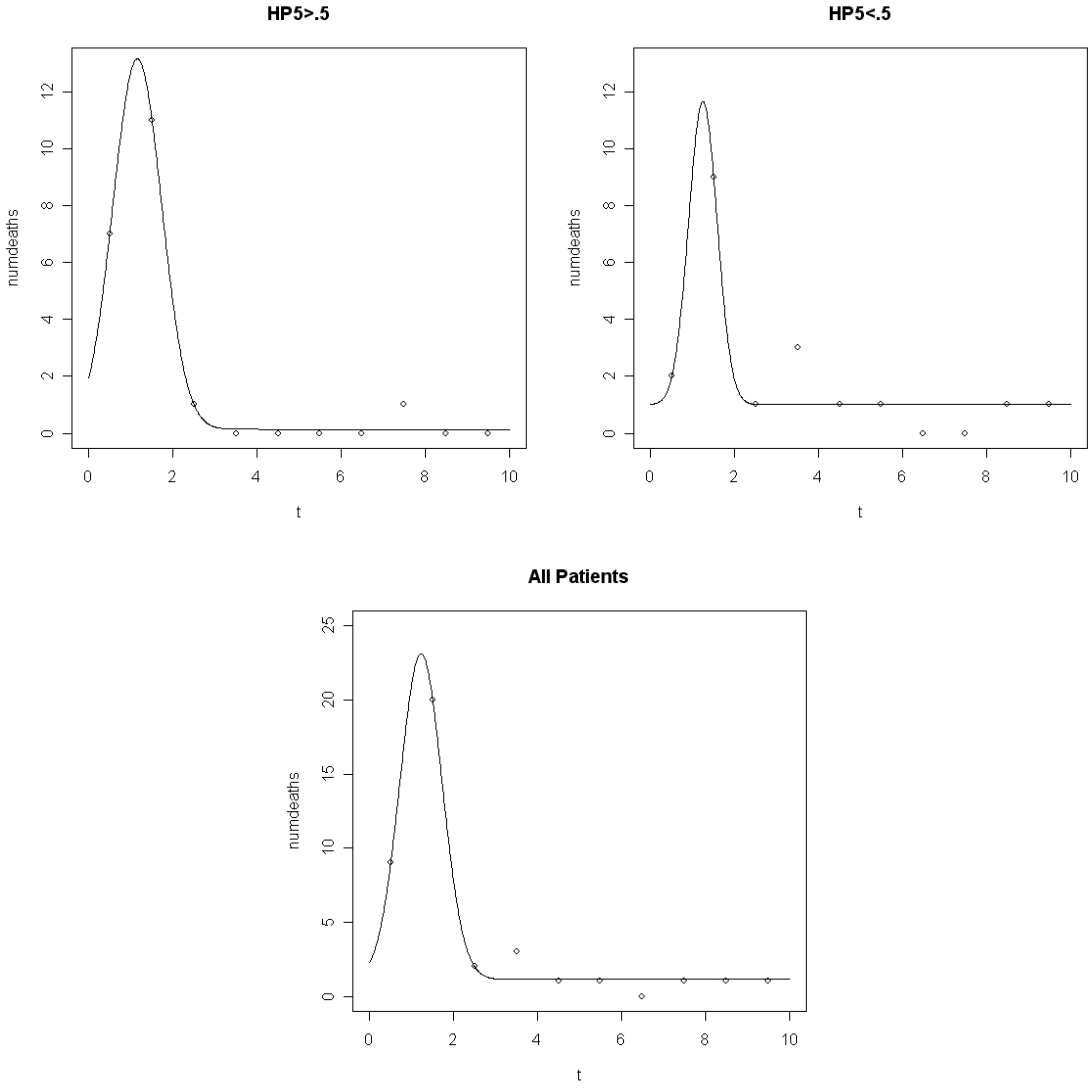


Figure 8.2: Plots of the various $\lambda(x)$ fitted against the data.

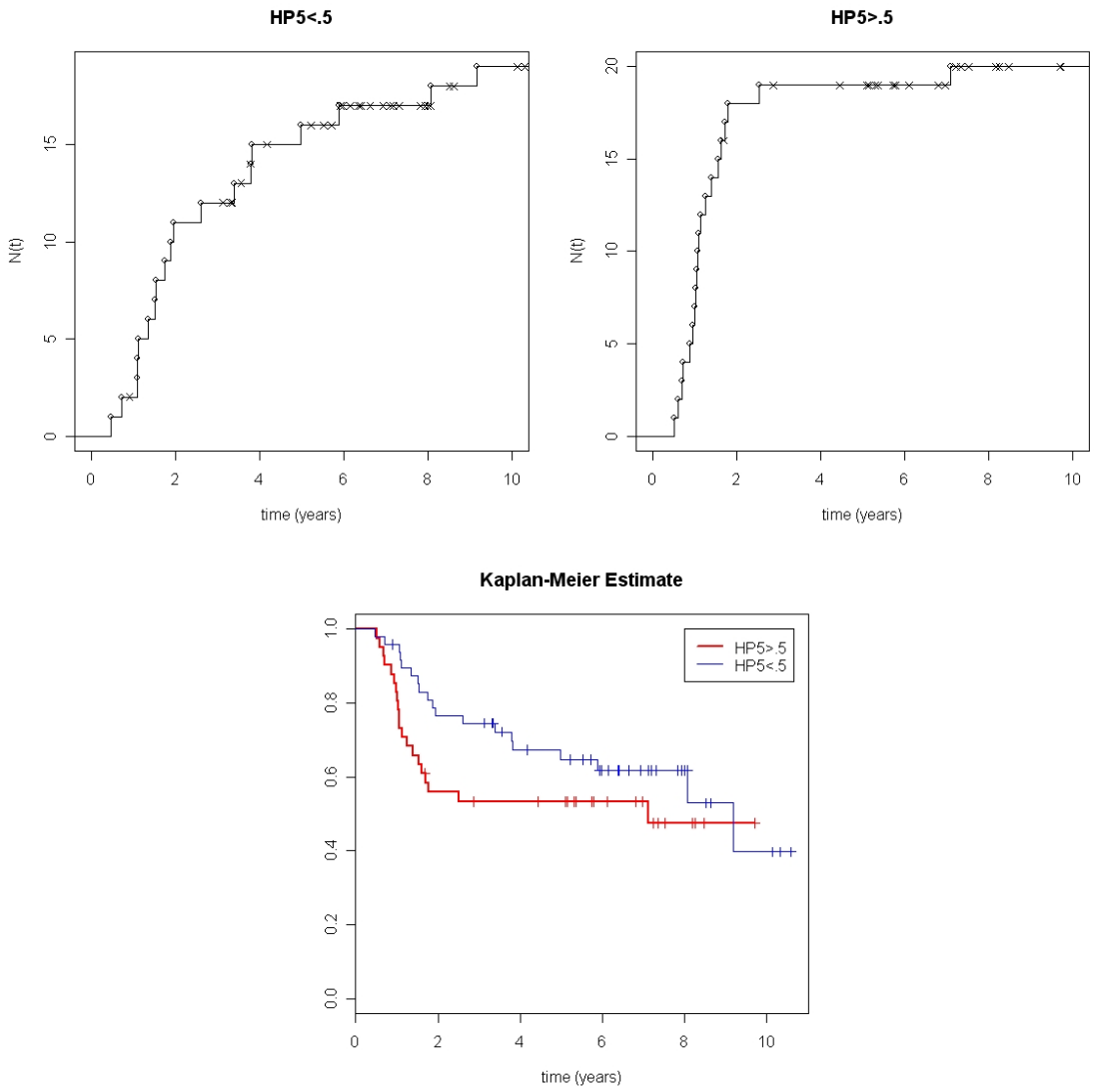


Figure 8.3: Plots of the survival data as an inhomogeneous Poisson process. The crosses indicate the time of last checkup for patients who are healthy. The bottom plot is the Kaplan-Meier curve of the same data as a comparison.

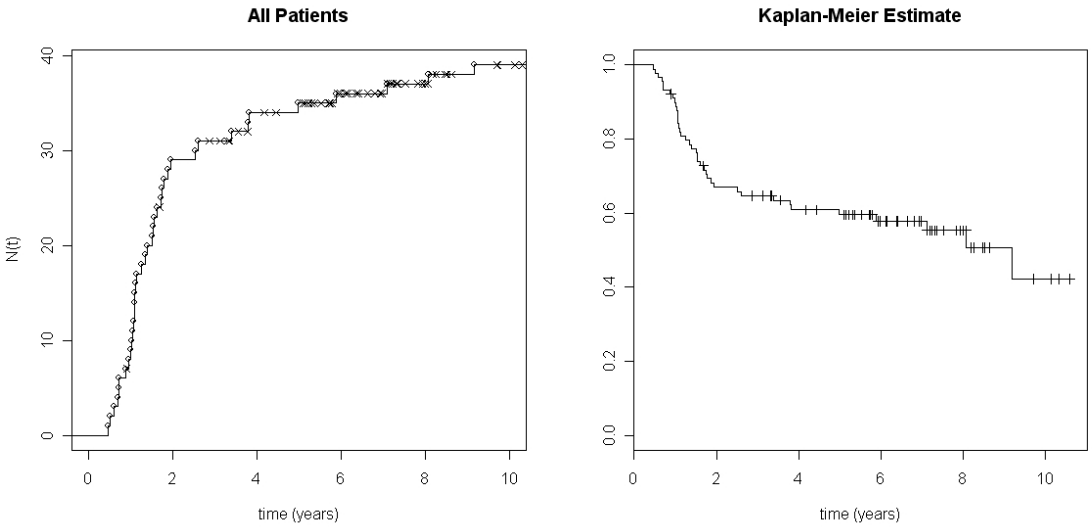


Figure 8.4: Survival plots for the data without stratification into two groups

Now lets consider some absolute probabilities at various times (figure 8.6). We can make several observations here; for example, in the table for $t = 1$, the most probable number of deaths for the hypoxic group is 7 whereas for the non-hypoxic group it is 3. So again we see that in the short term the hypoxic group is more likely to have a higher mortality rate. This pattern holds for $t = 3$ and $t = 5$ but from the previous analysis we would expect the probabilities to be about the same for large t. Another thing to notice is that as t increases, the probability of having more deaths increases for both groups, which makes sense from a physical point of view.

Perhaps of more interest is the probability of at least a certain number of deaths at a given time. We can easily calculate such probabilities using the values in the tables and the rules of probability:

$$P(N(t) \geq k) = 1 - P(N(t) < k) = 1 - \sum_{i=0}^{k-1} P(N(t) = i) \tag{8.8}$$

$$= 1 - \sum_{i=0}^{k-1} p_i(t) \tag{8.9}$$

For example, the probability of at least 15 deaths at $t = 5$ is 0.86 for the hypoxic group compared to 0.44 for the non-hypoxic group. However, at $t = 10$ the probabilities are 0.89 and 0.85 respectively, once again reaffirming the idea that

Λ(t)			
t	H	NH	C
1	7.22	3.17	10.08
2	17.39	10.95	28.09
3	18.87	12.06	30.89
4	19.02	13.06	32.04
5	19.16	14.06	33.18
6	19.30	15.06	34.32
7	19.44	16.06	35.46
8	19.58	17.06	36.60
9	19.72	18.06	37.74
10	19.86	19.06	38.88

Λ(t)/t			
t	H	NH	C
1	7.22	3.17	10.08
2	8.70	5.47	14.05
3	6.29	4.02	10.30
4	4.76	3.26	8.01
5	3.83	2.81	6.64
6	3.22	2.51	5.72
7	2.78	2.29	5.07
8	2.45	2.13	4.57
9	2.19	2.01	4.19
10	1.99	1.91	3.89

Figure 8.5: Table for the intensity function $\Lambda(t)$ and the average rate of occurrence $\Lambda(t)/t$.

p_i(t)	t=1		
i	H	NH	C
1	0.005	0.133	4.21E-04
2	0.019	0.211	0.002
3	0.046	0.223	0.007
4	0.083	0.177	0.018
5	0.120	0.112	0.036
6	0.144	0.059	0.061
7	0.149	0.027	0.088
8	0.134	0.011	0.111
9	0.107	0.004	0.124
10	0.078	0.001	0.125

p_i(t)	t=3		
i	H	NH	C
5	1.27E-04	0.012	9.00E-09
10	0.010	0.104	8.37E-06
15	0.010	0.073	0.001
20	0.086	0.010	0.010
25	0.032	4.03E-04	0.044
30	0.005	6.01E-06	0.072

p_i(t)	t=5		
i	H	NH	C
5	1.03E-04	0.004	1.31E-09
10	0.009	0.065	1.74E-06
15	0.063	0.099	1.94E-04
20	0.087	0.029	0.004
25	0.035	0.003	0.026
30	0.005	8.12E-05	0.062

Figure 8.6: Table of absolute probabilities $p_i(t)$ at various times.

the hypoxic patients do worse in the short term but are not too different from the non-hypoxic patients in the long term.

8.3 Conclusion

The use of the Poisson process as a tool for analyzing the survival data offers us a simple and straightforward way of generating results that are easy to interpret. As was the case in the difference between the original estimate of HP5 and the Bayesian estimate of HP5, the difference between the Kaplan-Meier estimator and the Poisson process is that the Kaplan-Meier estimator generates a number, where as the Poisson process gives as a probability distribution. From the probability distribution, we are able to generate a wealth of information that is not as readily available from the Kaplan-Meier estimate, and from which we can gleam more concrete results.

From the table of probabilities and average number of deaths $\Lambda(t)$ there is evidence that the hypoxic group has a lower chance of survival in the short term, but is not much different from the non-hypoxic group in the long term. Such an observation may be because of a lack of data (few deaths) after the first few years, or it may well be because that such is the case in real life, that somehow hypoxia is only an indicator of short term survival and its effects are diminished for some reason if you can live through it for a few years.

Chapter 9

Summary, Conclusion, and Future Direction

In this thesis I have outlined in great detail why tumour oxygen concentration is important and how readings from a probe that purportedly measures oxygen partial pressure in biological tissue is used by clinical researchers to predict survival. I have gone to great lengths to explain the difference between the Bayesian and Frequentist interpretation of probability, and the two main results of this thesis, the Bayesian estimates of HP5 and median of the pdf for the PO2 distribution, reflect those differences while still conforming to the existing work in the literature. It is my goal that these two improved estimates of tumour oxygenation eventually lead to better analyses of survival probability, but that will be future work.

Oxygen plays a large role in radiotherapy because it is believed that oxygen is necessary for radiation to permanently damage a tumour cell. In a process known as the oxygen fixation hypothesis, the indirect action of radiation creates free radicals within a tumour cell that can only bind with the DNA molecule if sufficient oxygen is present, and it is this binding that causes irreparable harm to the DNA, leading to cell death. If there isn't enough oxygen present, the damage will not be permanent, and the cancer cell will survive. Regions of low oxygen, hypoxic regions, naturally arise in tumours because of the chaotic and haphazard way in which it grows, resulting in abnormal and inefficient vascular networks. Thus, it is of interest to determine the oxygen concentration levels in a tumour as it can adversely affect radiotherapy.

To measure oxygen concentration with the tumour, a needle electrode known as

the Eppendorf probe is used. It is a probe that was originally used in chemistry for measuring oxygen concentrations in solutions and has now been adapted for use in humans. As we saw in chapters 3 and 4, there may be some serious issues with using the probe in biological tissue. In particular, we saw that the readings undergo large oscillations, even though physically they are being taken mere millimeters apart, and we saw also that there are impossible negative values for partial pressure, corroborating Albanese's [28] claim that it may not be possible to calibrate the probe correctly for biological tissue. However, despite these issues, the Eppendorf probe is still probably the best method available to measure oxygen concentration within tumours because it is the only device that is physically inserted inside the tumour. Hence, we need to develop methods of extracting the necessary information from the Eppendorf data.

Oxygen concentration within a tumour is usually quantified by a measure called HP5, the percentage of the Eppendorf measurements under 5 mmHg. I think that the abnormally large and negative values present in the Eppendrof data is one of the reasons why researchers use HP5; it conveniently allows them to avoid the issue of calibration, as the actual values of the measurements are not used, and allows them to condense the large number of measurements into an easily interpretable number. There is evidence to suggest that there is correlation between HP5 and survival, and we have shown how that evidence is usually presented in the form of the Kaplan-Meier estimator and log-rank test. However, the rejection of the null hypothesis in the log-rank test is often misinterpreted as evidence towards the researchers proposed alternate hypothesis, and, even worse, we were not able to reject the null hypothesis with our data, so there is much to be done in truly determining whether HP5 is correlated with survival probability.

Nonetheless, we have used Bayesian methods to estimate the HP5 because it is the most commonly used measure of tumour oxygen concentration. The original formulation is just a number, a percentage that seems to discard other potentially useful information from the Eppendorf data, and the Bayesian estimate of the HP5 is able to recover some of that information through the error bar. One of the main advantages of the Bayesian formulation of data analysis is that everything is represented by probability distributions, and by representing our estimate of the HP5 as a probability distribution we were able to determine how accurate we think our estimate is. Although the best estimate of the HP5 using the Bayesian formulation yields the exact same result as the original calculation, the error bar tells us how

much we can trust that estimate (as a function of sample size) and gives us added flexibility in the survival analysis. Moreover, the Bayesian formulation allows us to incorporate other information into the estimate and has predictive qualities as well (such as in the form of Laplace's rule of succession detailed in Appendix G).

There is no real mathematical or biological justification to choose HP5 as our measure of hypoxia, so a method of estimating the median PO2 of the underlying distribution using maximum entropy is also presented. The median value of the data was used as the measure before researchers invented the notion of hypoxic proportion, and it is still reported and compared to HP5 in the literature. Hence there is just as much reason to estimate the median PO2 as there is the HP5. It should be clear intuitively that the median value of our data is likely not to be the median value of the oxygen tension within the entire tumour, and the maximum entropy method outlined in chapter 7 gives us a way of estimating the true median value based on the data. The results demonstrate that our estimate of the true value can be markedly different from the sample median, and the multimodal pdfs seem to be a result of the heterogeneity inside a tumour.

We could of course develop methods to estimate any quantity we wish from the Eppendorf data, but the HP5 and median PO2 are the two most commonly used measures in the literature. The results presented in this thesis demonstrate how Bayesian methods can potentially improve our study of tumour oxygen concentration, and perhaps when we have developed better methods of assessing survival probability (that does not involve a p-value) we will be able to determine whether HP5 or median PO2 are actually good measures of tumour oxygenation, or whether we need to pursue some different quantity. The preliminary alternative using inhomogeneous Poisson processes in chapter 8 yields similar results to the Kaplan-Meier estimator but is a step in the right direction as we are able to gleam more potentially useful information in the pdfs that are generated.

APPENDICES

Appendix A

Henry's Law

A.1 Modified versions of Henry's Law

Henry's Law was first proposed by English chemist William Henry in 1803. Henry postulated that, at a certain fixed temperature, the amount of gas dissolved in a liquid is proportional to the partial pressure of the gas in equilibrium with the liquid (i.e. when the solution is saturated). Although this law has since been verified experimentally, and the proportionality constants measured and well established, modern chemists have found that this simple law only holds in limited situations and have since added modifications to the law [34].

The form of Henry's Law used by electro-chemists such as Fatt [24], Hudson [30], and Albanese [28] in deriving the current equation for a polarographic electrode is the aforementioned narrowly applicable version:

$$C = kP, \tag{A.1}$$

where C is the concentration of the dissolved gas in the solute, P is the partial pressure of the gas in equilibrium, and k is the Henry's Law constant. It is used without justification or qualification in their papers.

For a more modern discussion of Henry's Law, first let us define exactly what we mean by equilibrium. Consider a closed system where there are a number of components existing in different phases, after a long enough time, we expect the system to settle into a state where "the macroscopic properties of the phases will not change" [34]. In such a state, we must have that the temperature and pressure

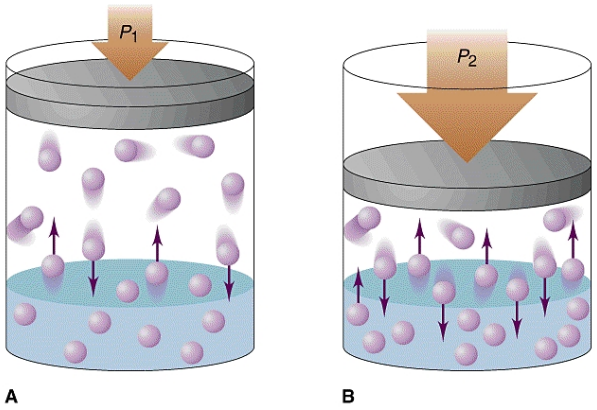


Figure A.1: Consider a closed container of a gas and a liquid under pressure. In equilibrium there will be a certain amount of the gas dissolved in the liquid. As we increase the pressure we expect that more of the gas will become dissolved because of the increased pressure on the system, which increases the partial pressure of the gas. Diagram from [38]

are equal for all phases, that the chemical potential, μ, must be the same for all components, and that the Gibbs free energy must be at a minimum. For a two phase system, say liquid (L) and vapour (V), this means that:

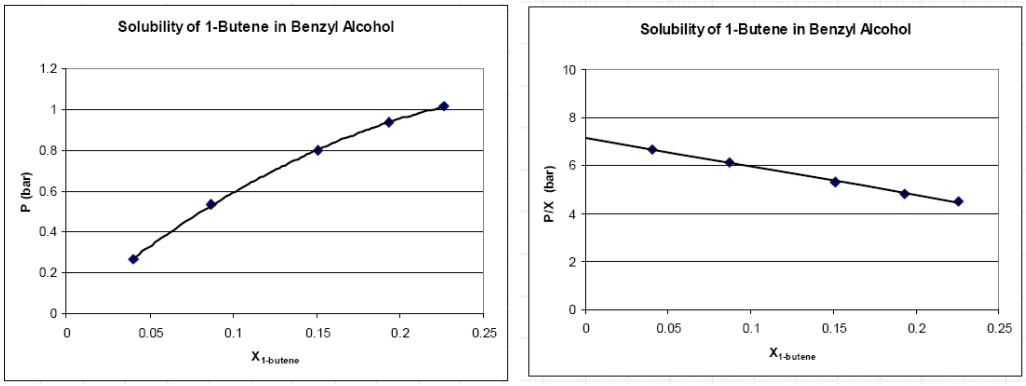
$$T^L = T^V, \tag{A.2}$$
$$P^L = P^V, \tag{A.3}$$
$$\mu_n^L = \mu_n^V, \quad \forall n, \tag{A.4}$$

where the subscript n denotes the different components of the system. We can rewrite the equation involving chemical potentials, (A.4), in terms of fugacity to get

$$f_n^L = f_n^V, \quad \forall n. \tag{A.5}$$

Fugacity is a thermodynamic quantity that acts as an adjusted pressure for non-ideal gases and is a measure of a substance's tendency to stay in one phase or another. It can be shown that fugacity is proportional to $\exp(\mu/\mathrm{RT})$ [39]. (More discussion on fugacity follows in the next section.) Given these equilibrium states, Henry's constant is defined as the limit

$$H_{ij} = \lim_{x_i \to 0} \frac{\hat{f}_i^L}{x_i}, \tag{A.6}$$



(a) Experimental data. The relationship between P and x seems to be non-linear.

(b) Plotting P/x gives a line that looks straight.

Figure A.2: To determine the Henry's Law constant k given a set of experimental data (left graph), it is convention to plot P/x against x and fit a straight line to the data (right graph). From the limit definition, k must then be the y intercept of the second plot. Figures from [37]

where H_{ij} denotes the Henry's Law constant for solute i in solvent j, $\hat{f}$ denotes the fugacity of a component in the solution, and x_i is the mole fraction of the solute in the liquid. Mole fraction is simply one way of measuring concentration, we can always choose another definition and adjust the dimensions and values of the constants to compensate, such choices have led to many different versions of Henry's Law and Henry's Law constants [35]. For example, in the case of an ideal gas, the fugacity is simply the pressure, and, using the usual definition of concentration, (A.6) simply reduces to the intuitive definition of Henry's constant

$$k = \lim_{C \to 0} \frac{P}{C}. \tag{A.7}$$

Figure A.2 demonstrates how one might compute this limit given some experimental data.

To derive a modified version of Henry's Law, consider the following thermodynamic relation that relates the fugacity, pressure, volume, and temperature of the liquid phase (a brief derivation is provided at the end of this section):

$$\frac{\partial \, ln[f_i^0]}{\partial P} = \frac{\overline{V}_i^{\infty}}{RT}, \tag{A.8}$$

where $\overline{V}_i^{\infty}$ is the molar volume (V/n) at infinite dilution (i.e. $x_i \to 0$) of component i, and f_i^0 is the reference fugacity of the liquid phase, related to the $\hat{f}_i^L$ by the

following equation:

$$\hat{f_i^L} = \gamma_i x_i f_i^0, \tag{A.9}$$

where γ_i is the activity coefficient. According to Carroll [34] as $x_i \to 0$ the activity coefficient approaches unity, so in the infinite dilution limit the reference fugacity f_i^0 is equal to the Henry's Law constant H_{ij}. Now if we integrate (A.8) from this infinite dilution limit ($x_i = 0, f_i^0 = H_{ij}, P = P_j^0$) to some arbitrary state, we get

$$f_i^0 = H_{ij} exp \left[\int_{P_j^0}^{P} \left(\frac{\overline{V}_i^{\infty}}{RT} \right) dP \right]. \tag{A.10}$$

The exponential term is known as the Poynting correction as it corrects for situations when the gas is not ideal. For the vapour phase, we have a relationship that relates the fugacity to the pressure:

$$\hat{f_i^V} = y_i \hat{\phi}_i P, \tag{A.11}$$

where y_i is the mole fraction of component i in the vapour, and $\hat{\phi}_i$ is the vapour phase fugacity coefficient which can be calculated (further details can be found in Carroll [34]). Recall that in equilibrium the fugacities must be equal, so we can combine equations (A.10), (A.9), and (A.11) to give

$$\gamma_i x_i H_{ij} exp \left[\int_{P_j^0}^{P} \left(\frac{\overline{V}_i^{\infty}}{RT} \right) dP \right]. = y_i \hat{\phi}_i P \tag{A.12}$$

This is known as the "Ensemble Henry's Law" and is the most generally applicable version of Henry's Law that relates pressure P to mole fraction x_i, but is difficult to use. In practise simplifying assumptions are made. For instance, if the component vapour pressure is low, or if the total pressure is low, then the Poynting term is approximately equal to unity, and if, in addition, the gas and liquid phases are close to ideal (practically, this means that the mole fractions x_i, y_i are small) then the activity and fugacity coefficients are also unity, $\gamma_i = 1$ and $\hat{\phi}_i = 1$, so the ensemble law simplifies dramatically to

$$x_i H_{ij} = y_i P. \tag{A.13}$$

Equation (A.13) is called "strict Henry's Law" and is applicable for total pressures up to about 2 atm (P<200kPa) and for liquid concentrations up to about about 1

mol% ($x_i < 0.01$). For a single solute where the solvent is nonvolatile ($y_j \simeq 0$, i.e. $y_i \simeq 1$) (A.13) further reduces to

$$x_i H_{ij} = P, \tag{A.14}$$

which we recognize as the original form of Henry's Law. For a far more extensive discussion on the various forms of Henry's Law the reader is referred to Carroll [34]

Both Carroll [35] and Smith and Harvey [36] outline many potential problems when using Henry's Law, the most important of which is the underlying temperature dependence of the Henry's Law constant that can dramatically change its value. Care must also be taken when looking up Henry's Law constants in a table, as many different measures of concentration and pressure can be used, changing the units and values of the Henry's Law constant.

A.2 Fugacity

Fugacity comes from the rearranged form of the 1st and 2nd laws of thermodynamics for Gibbs Free energy:

$$dG = -SdT + VdP. \tag{A.15}$$

Suppose that the temperature, T, is constant, then $dT = 0$. If we integrate (A.15) from some reference state (G_0, P_0), we get

$$G = G_0 + \int_{P_0}^{P} VdP. \tag{A.16}$$

Further, if the gas is ideal, we have $PV = nRT$, so this simplifies to

$$G = G_0 + nRTln\left(\frac{P}{P_0}\right). \tag{A.17}$$

Dividing both sides by n, we define G/n, the Gibbs Free energy per mole, as the chemical potential μ:

$$\mu = \mu_0 + RTln\left(\frac{P}{P_0}\right). \tag{A.18}$$

If the gas is not ideal, no such simplification occurs, and we have

$$\mu = \mu_0 + RT\int_{P_0}^{P} \overline{V}(P')dP', \tag{A.19}$$

where $\overline{V} = V/n$.

G. N. Lewis in 1908 proposed that we preserve the form of (A.18) by introducing a quantity called fugacity, $f(P)$, so that

$$\mu = \mu_0 + RTln\left(\frac{f(P)}{P_0}\right). \tag{A.20}$$

Physically, fugacity has the same units as pressure, and thus acts as an adjusted measure of pressure for non-ideal gases. As the pressure decreases, gases become ideal, so we require that $f(P)$ must satisfy the following condition:

$$\lim_{P\to 0} f(P) = P. \tag{A.21}$$

A.3 Derivation of equation A.8

To get equation (A.8), note that at constant temperature, we can divide (A.15) by n and differentiate (A.20) to get two different expressions for $d\mu$:

$$d\mu = \overline{V}dP, \tag{A.22}$$

$$d\mu = RT \cdot d\ lnf. \tag{A.23}$$

These expressions for $d\mu$ must be the same, thus

$$\begin{aligned} RT \cdot d\ lnf &= \overline{V}dP, \\ \Rightarrow \frac{d\ lnf}{dP} &= \frac{\overline{V}}{RT}. \end{aligned} \tag{A.24}$$

A more detailed discussion on fugacity can be found in the University of Arizona notes [39].

Appendix B

Derivation of the Spearman's Coefficient of Rank Correlation

One of the clearest expositions of this correlation coefficient is given in Sec. 12.3 of [49], and is summarized below (with a slight change in notation). Consider a random sample of n paired data

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n),$$

drawn from some bivariate population with (unknown) Pearson product-moment correlation coefficient

$$\rho(X, Y) = \frac{Cov(X, Y)}{[Var(X)Var(Y)]^{\frac{1}{2}}}, \tag{B.1}$$

where (X, Y) are the random variables from which the samples (x_i, y_i) are drawn. In orthodox statistics, it is usual for ρ to be estimated from the data using the sample correlation coefficient R, defined as

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}}, \tag{B.2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means for the $\{x_i\}$ and $\{y_i\}$ respectively. Now, suppose we rank the data in order of magnitude from smallest to largest (or vice versa, so long as we do the same to both) using the integers $1, 2, ..., n$, then we will have a new set of paired rank data

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n),$$

where

$$X_i = rank(x_i), \quad \text{and} \quad Y_i = rank(y_i). \tag{B.3}$$

For the Spearman coefficient, we do not use the actual data, but instead find the correspondence between the ranks. That is, we use our new set of data (X_i, Y_i) and plug them into equation B.2. The correlation between the ranks of the data is considered a measure of the correlation between the actual data, but it is not clear how exactly the two are related. Nonetheless, it is a commonly used measure because there is "intuitive appeal" [49] in considering just how the data are related to each other when ranked in order. I suppose it is like looking at a race, where we only care about who crosses the finish line first, and not really about how fast they did it, unless they have broken a world record or accomplished some other amazing feat.

Noting that the new data can only take on discrete values $1, 2, ..., n$, we can make several simplifications as follows:

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \tag{B.4}$$

$$\overline{X} = \overline{Y} = \frac{n+1}{2}, \tag{B.5}$$

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right)^2. \tag{B.6}$$

Substituting these results into equation 4.3 yields

$$R = \frac{12 \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n(n^2 - 1)}, \tag{B.7}$$

as the equation for the Spearman Rank Correlation Coefficient. It is often useful to write B.7 in terms of the differences instead. Let

$$D_i = X_i - Y_i = (X_i - \overline{X}) - (Y_i - \overline{Y}), \tag{B.8}$$

then we have that

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 + \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - 2 \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}), \tag{B.9}$$

and hence

$$R = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}. \tag{B.10}$$

This form of the Spearman coefficient is one most commonly used.

Appendix C

The Log-rank Test

Data in clinical trials are commonly split into two or more groups (e.g. control and experimental drug) and we want to see if there is any difference in survival between the two groups. The log-rank test is probably the most commonly used technique to examine the differences in survival between two (or more) groups and it does so by testing the null hypothesis that there is no difference in survival between the two (or more) groups [44][45][46].

For two groups of data, let us denote the failure times of all the patients as $t_{(1)} < t_{(2)} < ... < t_{(r)}$ as before. Now let d_{1j} be the number of deaths in group 1 at time $t_{(j)}$, n_{1j} the number of patients at risk in group 1 at $t_{(j)}$, and similarly d_{2j} and n_{2j} for group 2. Define $d_j = d_{1j} + d_{2j}$ and $n_j = n_{1j} + n_{2j}$ as the total number of deaths and total number at risk at time t_j. To consider the null hypothesis that there is no difference between the survival times of the two groups, we will compare the observed number of deaths in each group to the expected number of deaths under the null hypothesis. Notice that if we consider d_j, n_{1j}, and n_{2j} to be fixed, then the number of deaths in both groups can be specified by d_{1j} alone. So we will consider d_{1j} as a random variable that can take on any value in the range 0 to $\min(d_j, n_{1j})$. In fact, d_{1j} has a hypergeometric distribution [45]:

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}. \tag{C.1}$$

The expected value, or mean, of d_{1j} in this case is simply the intuitive ratio

$$e_{1j} = d_j \frac{n_{1j}}{n_j}. \tag{C.2}$$

Now we need to compute the difference between the observed number of deaths in group 1 and the expected number of deaths in group 1:

$$U = \sum_{j=1}^{r} (d_{1j} - e_{ij}) = \sum_{j=1}^{r} \left(d_{1j} - n_{1j} \frac{d_j}{n_j} \right). \tag{C.3}$$

Notice that this sum is simply the total number of deaths in group 1 minus the total number of expected deaths in group 1: $\sum d_{1j} - \sum e_{1j}$. Collett [45] claims that U can be shown to be normally distributed with zero mean, and that the variance is simply the sum of the variances of each d_{1j} since they are independent. Specifically, the estimated variance for U is given by

$$\widehat{Var}(U) = \sum_{j=1}^{r} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}, \tag{C.4}$$

and we have that

$$\frac{U}{\sqrt{\widehat{Var}(U)}} \sim N(0, 1). \tag{C.5}$$

We define the log-rank statistic as

$$S = \frac{U^2}{\widehat{Var}(U)}. \tag{C.6}$$

where $S \sim \chi^2$ with one degree of freedom. The p-value for this statistic is then the probability of getting a value greater than or equal to the value calculated from our data.

Appendix D

Derivation of the Binomial Distribution

To see why we must get the binomial distribution (and why we curiously denote the number of successes by r), following Sivia [2], consider an urn of W white balls and R red balls, and randomly draw one ball out of the urn. Suppose we have labeled the balls from $j = 1, ..., R + W$, then, assuming total ignorance as before, we would have that the probability of drawing the j^{th} ball is

$$P(j|I) = \frac{1}{R+W}, \tag{D.1}$$

since all balls are equally likely to be drawn. Thus the probability of drawing a red ball is, using the product rule and marginalization,

$$\begin{aligned} P(red|I) &= \sum_{j=1}^{R+W} P(red, j|I) = \sum_{j=1}^{R+W} P(red|j, I)P(j|I), \\ &= \frac{1}{R+W} \sum_{j=1}^{R+W} P(red|j, I), \end{aligned} \tag{D.2}$$

where $P(red|j, I)$ is the probability of the j^{th} ball being red *after* we have drawn it. So it is 1 if the ball is red, and 0 if it is white. This means that

$$P(red|I) = \frac{R}{R+W}. \tag{D.3}$$

This result should be no surprise, as we have just shown that probability theory agrees with our intuition that $P(red|I)$ should be

$$P(red|I) = \frac{\text{number of cases favourable to red}}{\text{total number of equally possible cases}}. \tag{D.4}$$

Now consider the case where we draw multiple times with replacement, what is the probability that in N trials we will have drawn r red balls? Using the product rule and marginalization again, we get

$$\begin{aligned} P(r|N,I) &= \sum_k P(r, S_k|N,I), \\ &= \sum_k P(r|S_k,N,I)P(S_k|N,I), \end{aligned} \tag{D.5}$$

where the sum is taken over all 2^N possible sequences of red-white drawings $\{S_k\}$ in N draws. $P(r|S_k,N,I)$ is, similar to before, either 1 when there are exactly r red balls and zero otherwise. $P(S_k|N,I)$ is the probability that a draw contains r exactly red balls. We will assume that the prior information I is simply the state of total ignorance, so all we know is that the draws are independent and do not influence the probability of subsequent draws. Hence the probability, $P(S_k|N,I)$, will not depend on the order and only on the number of each ball present

$$P(S_k|N,I) = [P(red|I)]^r \left[P(white|I)\right]^{N-r} = \frac{R^r W^{N-r}}{(R+W)^N}. \tag{D.6}$$

So the probability $P(S_k|N,I)$ can be taken out of the sum, and the sum itself is just the total number of possible combinations of r red balls in N draws with replacement, which is the binomial coefficient $\binom{N}{r}$. Thus, we get that

$$P(r|N,I) = \binom{N}{r} \frac{R^r W^{N-r}}{(R+W)^N}. \tag{D.7}$$

Now suppose that we were not thinking about urns, red balls, and white balls, but any kind of generic success or failure. We can rearrange D.7 in a more suggestive form. Let's define p as the probability of drawing a red ball, and q as the probability of drawing a white ball:

$$p = \frac{R}{R+W}, \qquad q = \frac{W}{R+W}. \tag{D.8}$$

Then $p+q=1$,and we have

$$P(r|N,I) = \binom{N}{r} p^r q^{N-r} = \binom{N}{r} p^r (1-p)^{N-r}, \tag{D.9}$$

which is precisely the binomial distribution. Note that p and q do not have to be the probabilities of getting red and white balls, they can represent red, and not-red, or more generically anything that we can simply label as success or failure, and the analysis will still hold.

Appendix E

Assigning Probabilities - The Maximum Entropy Principle

We have seen already that in the case where there are only two possible outcomes that it is natural to assign the binomial distribution for the probability. In general, we would like to develop methods for assigning probabilities based on our available information, whether it be for the prior, the likelihood function, or some other unknown quantity.

E.1 Principle of Indifference

The flat prior that we saw previously is normally called the *principle of insufficient reason* or the *principle of indifference*. Quoting Sivia [2], "it states that if we can enumerate a set of basic, mutually exclusive, possibilities, and have no reason to believe that any one of these is more likely to be true than another, then we should assign the same probability to all."Imagine we had a die, lets list the six possible outcomes as

$$X_i \equiv \text{ the top face has } i \text{ dots}, \tag{E.1}$$

for $i = 1, 2, .., 6$. Then, if we didn't know otherwise, we would simply assign

$$P(X_i|I) = \frac{1}{6}. \tag{E.2}$$

We must be very careful in interpreting this statement, it appears as if we have assumed the die to be fair, but that is not true. Imagine we had relabeled the propositions X_i as letters A,B,..,F, in some way, say X_3 was A and X_6 as B and

so on. After prescribing the probabilities $P(A|I)$, $P(B|I)$, ..., $P(F|I)$, suppose we were told that there was some mistake and A actually corresponded to X_4 and B to X_2. Since we have assumed that we have no knowledge of the situation, such a reordering should not change the probabilities because if it did it would mean that we have acquired other information that we didn't have before. Thus, the principle of consistency requires that we do not change the probability assignments, and the only way to ensure that this is true is by assigning the same probability to each possible outcome.

E.2 Location and Scale Parameters

The argument in the previous section worked for discrete outcomes and we can generalize it to the case of continuous parameters. Suppose X represented the position of some object, say the redial distance of the moon from the surface of the earth, then the probability that X lies in the range x to $x + \delta x$ is given by

$$P(X = x|I)dX = \lim_{\delta x \to 0} P(x \leq X < x + \delta x|I). \tag{E.3}$$

Similar to the argument before, suppose we were told that our instruments were wrongly calibrated and all the positions were actually offset by some constant x_0, if we were totally ignorant about the situation, such an adjustment should not make much difference to our posterior probability (other than possibly throwing some values out of our range of allowed values), so we need that

$$P(X|I)dX \simeq P(X + x_0|I)d(X + x_0). \tag{E.4}$$

Since $d(X + x_0) = dX$, this statement just reduces to $P(X|I) \simeq P(X + x_0)$, or in words

$$P(X|I) \simeq \begin{cases} \text{constant in the allowed range,} \\ 0 \quad \text{otherwise.} \end{cases} \tag{E.5}$$

So for location parameters, complete ignorance is captured by the assignment of a constant prior probability.

For quantities where the relative or fractional change is more important, rather than the absolute value as in the case of location or distance, we would need our pdf to be invariant to a change in scales. For instance, if were interested in the length L of some quantity, and were told that the measurements were actually in

millimeters instead of centimeters, no knowledge about the length scale of L should have much of an effect on our assignment of the probability. Such a requirement would mean that

$$P(L|I)dL \simeq P(\beta L|I)d(\beta L), \tag{E.6}$$

where β is a positive constant. Since $d(\beta L) = \beta dL$, the only possible assignment is

$$P(L|I) \propto \begin{cases} \frac{1}{L} & \text{in the allowed range,} \\ 0 & \text{otherwise.} \end{cases} \tag{E.7}$$

This probability is known as Jeffreys' Prior, and expresses our ignorance about scale parameters.

E.3 The Maximum Entropy Principle

E.3.1 Testable Information

For more complicated situations, where we actually have some information about the quantity of interest, we need more sophisticated methods. According to Sivia, to make the problem more manageable, Jaynes in 1957 demanded that our prior information I be *testable*. What do we mean by testable? Imagine that we were in possession of a die again, but this time we were told that over a large number of trials, the die on average gave a result of 4.5. We can write that piece of information mathematically as

$$< X >= \sum_{i=1}^{6} iP(X_i|I) = 4.5. \tag{E.8}$$

Given such a concrete restriction, we can easily test candidate pdfs to see whether they satisfy the constraint. For instance, the uniform pdf that we assigned previously would not satisfy the constraint, as its expected value is 3.5. Thus, we call the constraint testable because we can test candidate pdfs for their eligibility as a possible pdf under the information given. Things that are not testable would be vague statements along the lines of "I believe the die is biased" or "it seems to roll a six more often than the other numbers".

E.3.2 Entropy

If our prior information is testable, Jaynes goes on to propose that the best choice for the probability is one that maximizes the **entropy** S while satisfying all the

constraints. For the die example, abbreviating by letting $p_i = P(X_i|I)$, we would have

$$S = -\sum_{i=1}^{6} p_i ln(p_i), \tag{E.9}$$

subject to the expected value constraint and normalization:

$$\sum_{i=1}^{6} ip_i = 4.5, \quad \text{and} \quad \sum_{i=1}^{6} p_i = 1. \tag{E.10}$$

To solve, we would use the method of lagrange multipliers to get

$$Q = -\sum_{i=1}^{6} p_i ln(p_i) + \mu \left(1 - \sum_{i=1}^{6} p_i\right) + \lambda \left(4.5 - \sum_{i=1}^{6} ip_i\right), \tag{E.11}$$

where μ and λ are our lagrange multipliers. After some work, we get that

$$p_i = e^{-(\mu+1)} e^{-\lambda i}, \tag{E.12}$$

where the constants will be determined by the constraints.

A natural question that one would ask is: why maximize this odd looking quantity S? The actual expression for entropy comes from statistical mechanics and information theory, and there are various justifications as to why it is desirable to maximize it in the setting of probability theory [1][2]. We will look at one heuristic justification known as the monkey argument.

E.3.3 The Monkey Argument

Suppose there are M distinct possibilities $\{X_i\}$ and some given criterion expressible as testable information in our background information I. What is the best way to assign the probabilities $p_i = P(X_i|I)$? By best, we mean that we want the constraint to be satisfied while assuming as little else about the p_i as possible. To answer this question, imagine that we had replaced the propositions X_i with M boxes of the same size, and we had a team of monkeys randomly throwing pennies at the boxes. We will assume that the monkeys have no particular preference to throw the pennies at any one box so that after a large number of tosses, we will have a candidate distribution for p_i. If the distribution does not satisfy the constraint in I we discard it and start over, if it does, we make note of it and begin another trial. Repeating the trials many times will yield a large number of candidate distributions

that satisfy I, but some will reappear more often than others, and the one that is repeated the most would be our best choice for the distribution. It is the best choice because we have assumed that the monkeys have no preferred disposition to throw in one direction over another, so the resulting most likely distribution should be the one that sanctifies the testable constraint while being least committal about anything else.

To see why these monkeys give us the expression $-\sum p_i ln(p_i)$, denote the number of coins in box i by n_i, and let the total number of pennies thrown be N, so that

$$N = \sum_{i=1}^{M} n_i. \tag{E.13}$$

We will assume that N is very large, in particular, it must be much larger than the number of boxes $N \gg M$. The distribution of the $\{n_i\}$ gives us a candidate distribution for the $\{p_i\}$:

$$p_i = \frac{n_i}{N}, \tag{E.14}$$

where $i = 1, 2, ..., M$. Since the N pennies are equally likely to land in any of the M boxes, there are M^N ways of distributing the coins, but not all of these possibilities are distinct (i.e. yield the same $\{n_i\}$), so the expected frequency F of obtaining $\{p_i\}$ is

$$F\left(\{p_i\}\right) = \frac{\text{number of ways of obtaining } \{n_i\}}{M^N}. \tag{E.15}$$

The numerator is simply the multinomial coefficient:

$$\frac{N!}{n_1! n_2! ... n_M!}. \tag{E.16}$$

So we have that

$$ln[F] = -N ln[M] + ln[N!] - \sum_{i=1}^{M} ln[n_i!]. \tag{E.17}$$

The reason we take logarithms here is similar to the reason why we took logarithms in the derivation for the error bar: we want to estimate the quantity F, and taking the logarithm makes the function vary more slowly. Using Stirling's approximation

$$ln[n!] \simeq n ln[n] - n, \tag{E.18}$$

which is valid for large n, we have that

$$ln[F] = -N ln[M] + N ln[N] - \sum_{i=1}^{M} n_i ln[n_i]. \tag{E.19}$$

Substituting for n_i using p_i and noting that $\sum p_i = 1$, gives

$$ln[F] = -Nln[M] - N\sum_{i=1}^{M} p_i ln[p_i]. \tag{E.20}$$

Since F is the frequency for which a particular candidate distribution $\{p_i\}$ appears, and since ln is monotonic, we can find the most likely distribution, the one that maximizes F, by maximizing lnF. Therefore, we want to maximize the quantity:

$$S = -\sum_{i=1}^{M} p_i ln[p_i], \tag{E.21}$$

which is exactly the entropy expression we had before. Note that we use the natural logarithm here for convenience, it could be log to any base, or even another monotonic function. So the final form of the entropy for this monkey example is simply a choice that allows us to maximize the frequency of occurrence, F, to obtain the most likely probability distribution $\{p_i\}$. There are more rigorous arguments as to why the particular form $-\sum p_i log[p_i]$ is required or desirable, such as Shannon's theorem on information theory, reproduced in Jaynes [1] on pages 346-51.

E.3.4 Extending to continuous parameters

Suppose that the boxes in the previous example were not of equal sizes. Let the probability that a monkey will throw a penny into box i be m_i, then we have that

$$\sum_{i=1}^{M} m_i = 1. \tag{E.22}$$

This is just the generalization of the previous case where the m_i were all equal to $1/M$. Now the frequency F that there are n_i coins in the ith box is the multinomial distribution:

$$F(\{p_i\}) = \frac{N!}{n_1!n_2!...n_M!} m_1^{n_1} m_2^{n_2} ... m_M^{n_M}. \tag{E.23}$$

Taking logarithms as before, we get:

$$ln[F] = \sum_{i=1}^{M} n_i ln[m_i] - N\sum_{i=1}^{M} p_i ln[p_i]. \tag{E.24}$$

Substituting $n_i = p_i N$ then yields

$$\frac{1}{N} ln[F] = -\sum_{i=1}^{M} p_i ln\left[\frac{p_i}{m_i}\right] =: S. \tag{E.25}$$

This extension to boxes of different sizes is necessary if we want to extend the formulation to continuous variables:

$$S = -\int p(x) ln \left[\frac{p(x)}{m(x)}\right] dx. \tag{E.26}$$

The reason is that want S to be invariant under a change of variables $x \rightarrow y = f(x)$, and choosing an appropriate $m(x)$ that changes in the same way as $p(x)$ will ensure that occurs. The situation is very similar to the one discussed previously regarding scale and location parameters.

Interpretation of $m(x)$

To see what kind of role $m(x)$ plays, consider the situation where we are totally ignorant about $p(x)$. In that case, all we know is that the probability must be normalizable: $\int p(x)dx = 1$. Considering the discrete case instead to simplify the algebra, we want to maximize E.25 subject to the condition $\sum_i p_i = 1$, and we can do that by using the method of Lagrange multipliers. Let

$$Q = -\sum_i p_i ln \left[\frac{p_i}{m_i}\right] + \lambda \left(1 - \sum_i p_i\right), \tag{E.27}$$

assuming the $\{p_i\}$ are independent, we have that

$$\frac{\partial Q}{\partial p_j} = -1 - ln \left[\frac{p_j}{m_j}\right] - \lambda = 0, \tag{E.28}$$

since $\partial p_i / \partial p_j = 0$ for $i \neq j$. A simple rearrangement produces

$$p_j = m_j e^{-(1+\lambda)}. \tag{E.29}$$

Recall that λ is simply the lagrange multiplier, a constant, so we have established that

$$p(x|\text{normalization}) \propto m(x). \tag{E.30}$$

Thus $m(x)$ is a multiple of the probability distribution that expresses the ignorance of the situation at hand. Since $m(x)$ itself is a probability, we would need to assign it using the same tricks described earlier in the chapter, and we normally call it the *measure* of ignorance since it reflects how much we know about the situation a priori.

E.3.5 Some examples

Here are some examples of how common pieces of testable information that we might have access to will lead to some recognizable distributions using maximum entropy. These examples illustrate how using the principle of maximum entropy actually agrees with our logic and intuition.

Mean $\Rightarrow$ Exponential Distribution

Suppose we knew that the expected value of our unknown probability distribution had to be some fixed number μ. Then we would have the constraint

$$< x >= \int xP(x|I)dx = \mu. \tag{E.31}$$

For simplicity, consider again the discrete case instead. We want our probability distribution to be normalized and give the correct expected value, so the function we want to maximize is the following:

$$Q = -\sum_i p_i ln\left[\frac{p_i}{m_i}\right] + \lambda_0\left(1 - \sum_i p_i\right) + \lambda_1\left(\mu - \sum_i x_i p_i\right), \tag{E.32}$$

where the λ_i are the lagrange multipliers. Setting $\partial Q/\partial p_j = 0$ and solving as before yields

$$p_j = m_j e^{-(1+\lambda_0)} e^{-\lambda_1 x_j}. \tag{E.33}$$

In the discrete case, the lagrange multipliers would normally be calculated numerically by applying the two constraints, assigning equal values for m_j given that our background knowledge does not tell us anything about how to assign the probabilities, and introducing the partition function $Z = \sum exp(-\lambda_1 x_j)$ to simplify the algebra. For the continuous case, we will similarly assign a uniform measure for $m(x)$ and conclude from the discrete version that

$$P(x|I) \propto e^{-\lambda_1 x}. \tag{E.34}$$

If the admissible range for x is $[0, \infty)$ then we can determine the normalization constant and λ_1 analytically to give

$$P(x|\mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \tag{E.35}$$

which is commonly known as the exponential distribution. Choosing other measures for $m(x)$ will of course give rise to different distributions.

Mean + Variance ⇒ Gaussian Distribution

Now imagine in addition to the mean μ, we also had the variance σ^2:

$$< (x-\mu)^2 >= \int (x-\mu)^2 P(x|I)dx = \sigma^2. \tag{E.36}$$

Note that here we have in fact specified two pieces of information, $< x >= \mu$ and $< x^2 >$, and combined them succinctly into one expression using the relation $< (x-\mu)^2 >=< x^2 > -\mu^2$. Turning to the discrete case once again, we would have that

$$Q = -\sum_i p_i ln\left[\frac{p_i}{m_i}\right] + \lambda_0\left(1 - \sum_i p_i\right) + \lambda_1\left(\sigma^2 - \sum_i (x_i-\mu)^2 p_i\right), \tag{E.37}$$

and upon differentiating get

$$p_j = m_j e^{-(1+\lambda_0)} e^{-\lambda_1 (x_j-\mu)^2}. \tag{E.38}$$

Thus in the continuous case, applying a uniform measure for $m(x)$ for the same reasons as before, we have

$$P(x|I) \propto e^{-\lambda_1 (x-\mu)^2}. \tag{E.39}$$

If the limits of integration are $(-\infty, \infty)$, then, after a little bit of work, using the known result that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}, \tag{E.40}$$

we get the gaussian distribution

$$P(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{E.41}$$

This result should be of no surprise as the mean and variance fully characterize the normal distribution, but what we have shown here is actually almost the reverse: In the situation where only the mean and variance of some quantity is known, according to the principle of maximum entropy, the distribution that satisfies these constraints while being the least committal about anything else is the Gaussian.

Binomial Distribution

To derive the binomial distribution using maxent, suppose we knew that the expected number of successes in M trials was $< N >= \mu$. Then our constraint would be

$$< N >= \sum_{N=0}^{M} N P(N|M, \mu) = \mu, \tag{E.42}$$

and the normalization condition. Setting up the lagrange multipliers as before, we get the result

$$P(N|M,\mu) \propto m(N)e^{-\lambda N}, \tag{E.43}$$

where λ is the lagrange multiplier and $m(N)$ is the measure. Recall that $m(N)$ reflects our ignorance of the situation, so given only that we have N successes in M trials, we must assign equal probability to all 2^M possible outcomes. That means $m(N)$ must be the number of ways to obtain N successes in M trials:

$$m(N) = \binom{M}{N} = \frac{M!}{N!(M-N)!}. \tag{E.44}$$

To impose the normalization constant, recall the binomial theorem:

$$(a+b)^N = \sum_{j=0}^{N} \binom{N}{j} a^j b^{N-j}. \tag{E.45}$$

Since we want the sum of the probabilities to equal one, we have that

$$\sum_{N=0}^{M} m(N)e^{-\lambda N} = \left(e^{-\lambda}+1\right)^M. \tag{E.46}$$

This means the proportionality constant in E.43 is $\left(e^{-\lambda}+1\right)^{-M}$. To apply the constraint on the mean, differentiate E.46 with respect to λ to obtain

$$\sum_{N=0}^{M} Nm(N)e^{-\lambda N} = M\left(e^{-\lambda}+1\right)^{M-1} e^{-\lambda}, \tag{E.47}$$

and deduce that

$$\mu = M\left(1+e^{-\lambda}\right)^{-1}. \tag{E.48}$$

Thus, with a little manipulation, we arrive at the binomial distribution as expected:

$$P(N|M,\mu) = \binom{M}{N}\left(\frac{\mu}{M}\right)^N \left(1-\frac{\mu}{M}\right)^{M-N}. \tag{E.49}$$

Notice that the information we are given here to obtain the binomial distribution is identical to the information we were given when we considered the red and white balls in the urn, just phrased slightly differently. The expected number of red balls drawn in an urn that has R red balls and W white balls after $M = W+R$ draws has to be R, the difference here is that we have generalized the problem. The agreement between the maximum entropy method and logical deduction should give us faith that this strange quantity actually works.

Counting ⇒ Poisson Distribution

Remember in survival analysis we were interested in knowing whether a patient suffered a relapse, or worse, died, and noting when it happened. If we split continuous time into many tiny intervals, small enough so that not more than one event occurs in each interval, then we can actually consider each interval as an independent Bernoulli trial and consider the event, relapse or death, as a success. In general, for any situation where we have to count the occurrence of a certain event over some time interval, we can split the interval into M subintervals, let $M \to \infty$ so that no more than one event occurs in each subinterval, and consider each subinterval as an independent Bernoulli trial.

To that end, consider the limit $M \to \infty$ in the previous example. The initial information given, $< N >= \mu$, is the same so we arrive at the same initial result as E.43, but now N signifies the number of events observed over the time interval. For M large, we have that

$$\frac{M!}{(M-N)!} \simeq M^N, \tag{E.50}$$

which means our measure is now

$$m(N) = \frac{M^N}{N!}, \tag{E.51}$$

and hence the probability is

$$P(N|\mu) = A\frac{\left(Me^{-\lambda}\right)^N}{N!}, \tag{E.52}$$

where A is the normalization constant. To determine A, let's first write out the normalization condition:

$$\sum_{N=0}^{\infty} A\frac{\left(Me^{-\lambda}\right)^N}{N!} = 1. \tag{E.53}$$

Notice that the left hand side is actually the Taylor series for e^x, so the normalization condition reduces to

$$A = \frac{1}{e^{Me^{-\lambda}}}. \tag{E.54}$$

To apply the condition $< N >= \mu$, we would differentiate the series as before, and get that $Me^{-\lambda} = \mu$. Putting everything together gives us the Poisson distribution:

$$P(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}. \tag{E.55}$$

What is interesting here is that the final result is independent of M, the number of subintervals. This might lead one to believe that the size of M is not relevant, but keep in mind that we assumed M to be large right at the beginning.

E.3.6 Priors Vs Likelihood Functions

The maximum entropy method that we have developed for assigning probability distributions based on our prior information I can be used for both the prior and the likelihood function, and is simply a matter of nomenclature and personal preference. If the data is new, then we call the pdf assigned using this new data the likelihood function; if the data is old, then we call it the prior.

Appendix F

Preliminary Work - Filtering the Data

Prior to the adoption of Bayesian methodology to analyze the data, my initial work focused on using mathematical filters to filter the data in an attempt to improve the HP5 estimate. The results are presented in this appendix.

F.1 A few simple filters

Idea: Using the track information, we can plot the oxygen tension against position and consider each track as a sort of signal, allowing us to apply some signal processing filters to try and eliminate the noise.

F.1.1 Moving Average Filter

The moving average filter is a very simple filter that smooths a signal in the time domain, has low pass characteristics, and is, in some ways, ideal for filtering white noise [40]. In words, given an input signal $\{x[k]\}$, we generate a new sequence $\{y[k]\}$ by taking the average of $x[k]$ and several of its nearest neighbours. It is called a moving average filter because we can think of the filter as a moving window over the signal that computes the average of the signal under the window. (Illustrated in figure F.1.)

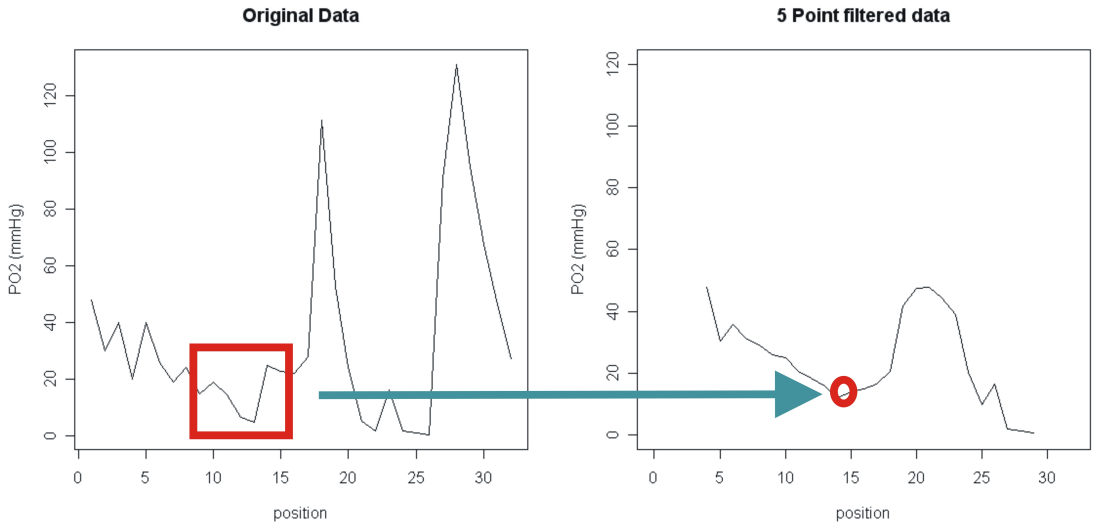


Figure F.1: A cartoon showing how a moving average filter works. We can think of the red box on the left as moving along the curve, taking the average of the values under it, and outputting values for the plot on the right

Mathematically, for an input $x[k]$ and a window size $n = 2N + 1$, the output $y[k]$ of the symmetric moving average filter is given by

$$y[k] = \frac{1}{2N+1} \sum_{i=-N}^{N} x[k+i]. \tag{F.1}$$

Note that we can implement this recursively as

$$y[k] = \frac{(2N+1)y[k-1] - x[k-N-1] + x[k+N]}{2N+1}. \tag{F.2}$$

For the end points we can either modify the scheme to a one sided average at those points or discard the points and have $y[k]$ be a shorter sequence. I have chosen to discard the ends since the length of the signal does not matter too much in calculating HP5. Examples using this filter are shown in figure F.2.

F.1.2 Threshold Moving Average Filter

A useful modification to the moving average filter is to add a threshold. Given a threshold value T, the threshold moving average filter is given by:

$$y[k] = \begin{cases} \frac{1}{2N+1} \sum_{i=-N}^{N} x[k+i], & \text{if } \Delta > T, \\ x[k], & \text{if } \Delta < T, \end{cases} \tag{F.3}$$

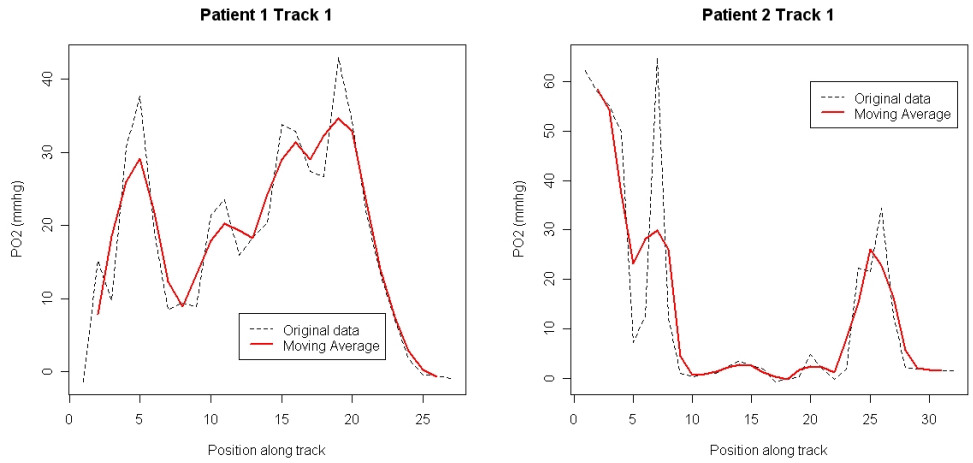


Figure F.2: Plots of the data passed through a 3 point moving average filter. The peaks and valleys in the signal are smoothed by the averaging of the values, and the ends are cut off by choice.

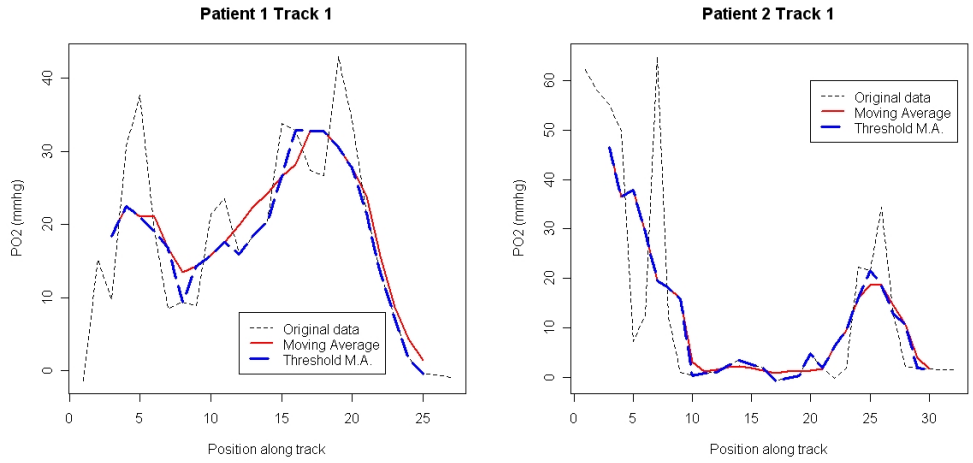


Figure F.3: 5 point threshold moving average with T=5. Notice how with the addition of the threshold, the output signal will follow the original signal if the original signal is not undergoing large oscillations.

where Δ is simply the difference

$$\Delta = \left| x[k] - \frac{1}{2N+1} \sum_{i=-N}^{N} x[k+i] \right|. \tag{F.4}$$

Again, I have chosen to cut of the ends of the signal where this symmetric definition is not valid.

The effect of the threshold is to not smooth the regions where there is little variation and preserve the original input signal in our output for those regions. Physically, regions of the signal where there is little variation are possible indicators of places that are not affected as much by noise, so retaining the original signal makes sense because the smoothing action of the moving average filter might cause information to be lost in regions where there is little noise. Use of this filter is illustrated in figure F.3.

F.1.3 Weighted Moving Average Filter

We can also add weights to each value in the input $\{x[k]\}$ to get the weighted moving average filter. Given some weights $(a_{-N}, a_{-N+1}, ..., a_N)$, the weighted moving average filter is given by

$$y[k] = \frac{\sum_{i=-N}^{N} a_i x[k+i]}{\sum_{i=-N}^{N} a_i}. \tag{F.5}$$

Although there is no particular reason (such as prior information) to weight each term in $\{x[k]\}$ differently, we might want to weigh the central term $x[k]$ more for the output $y[k]$ because it makes sense that the middle term should have more influence on the measurement at that point than the neighbouring terms. Some plots using the weighted moving average filter with different weights are demonstrated in figure F.4.

There are many more variations of the basic moving average filter, but the three aforementioned filters serve as a good basis to work with, as they are all simple to implement, make sense physically, and generate some interesting results.

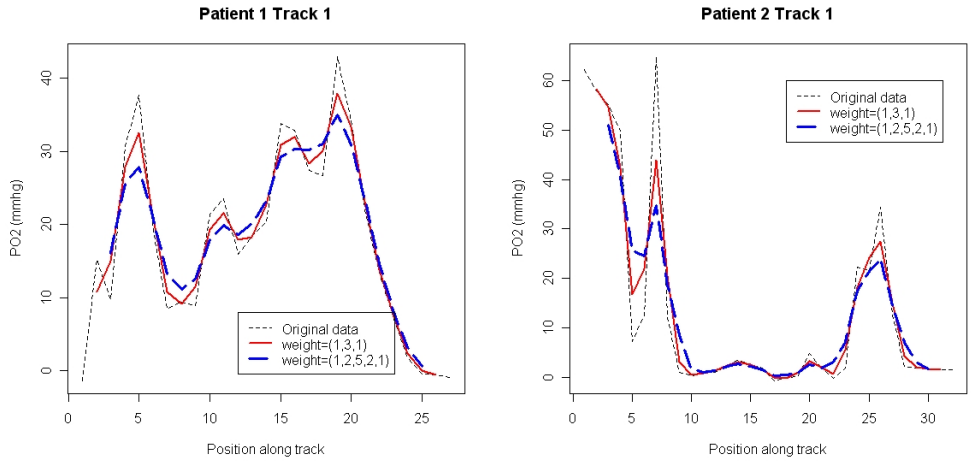


Figure F.4: Two kinds of weighted moving average filter. Centered weights make sense since the current value should have the most influence on the output signal.

F.2 Results

Recall that to classify the amount of hypoxia in a patient we calculate the HP5. Then, based on their HP5, the patients are split into two groups, oxic (HP5<.5) and hypoxic (HP5>.5), and put into the log-rank test to see whether there is any difference in survival between the two groups. Applying the various moving average filters to the data affects the HP5 classification on some patients, and that in turn affects the Kaplan-Meier plots and results of the log-rank test.

F.2.1 Effects of the filter on HP5

Applying the filters to the data and recalculating the HP5 changes the results. In particular, it changes the classification of some patients and that will affect the result of the log-rank test. For comparison, the table in figure F.5 displays some of the results for three different 5 point filters, moving average, threshold moving average with T=1, and weighted moving average with weights (1,2,5,2,1), and how they deviated from the original data. Around 10 patients changed classification in each case, and perhaps of interest is the fact that all of those who changed went from being classified as hypoxic to being classified as oxic after the filtering, which is probably a result of the filter eliminating small values in the data when smoothing over regions with both large and small PO2 values.

Patient Number	Original HP5	Moving Average	Threshold M.A.	Weighted M.A.	Changed classification?
1	0.520	0.510	0.427	0.451	N,Y,Y
2	0.438	0.328	0.313	0.330	N,N,N
3	0.568	0.432	0.345	0.379	Y,Y,Y
4	0.468	0.435	0.340	0.360	N,N,N
6	0.917	0.880	0.880	0.906	N,N,N
10	0.825	0.800	0.788	0.779	N,N,N
12	0.578	0.500	0.527	0.568	N,N,N
15	0.123	0.105	0.067	0.089	N,N,N
20	0.549	0.504	0.487	0.513	N,Y,N

Figure F.5: Some values of HP5 before and after filtering. The letters in the right most column, Y and N, stand for yes and no respectively to the question of whether any of the filters changed the classification of the patient. Notice that for the most part the filter does not change the classification unless the patient was borderline to begin with, such as for patients 1, 3, 20, and almost for 12.

F.2.2 Effects of the filter on the Kaplan-Meier plot and log-rank test

We can apply the same Kaplan-Meier estimator and log-rank test to the filtered data, and, as you can see in the table in figure F.7, the filtered data did not have much of an effect on the p-value. However, recall that one of the reasons 50% was chosen as the threshold value for HP5 for splitting the patients into two groups was because it was the median value of HP5 in the unfiltered data, so it might not make sense to keep 50% as the threshold for the filtered data. If the groups become imbalanced after the filtering, we might simply choose the median or some other value of HP5 to split the patients into two new roughly equally sized groups instead. We can see most clearly how keeping 50% might be a bad choice for the filtered data when looking at the plots of the Kaplan-Meier estimator (figure F.6) and p-value for the threshold moving average filter (figure F.7). Using 50% as the threshold for the aforementioned data yields a p-value of 0.473, much larger than the original value of 0.063, but when the median value of 41% was used for splitting, the p-value for the filtered data was very close to the original value (0.062). Testing some arbitrary threshold choices between 40% and 50% gave 42% as a threshold value that "improved" the log-rank statistic (lowered it down to 0.03). There is no mathematical or physical reason to choose 42% and that number was simply found using trial and error.

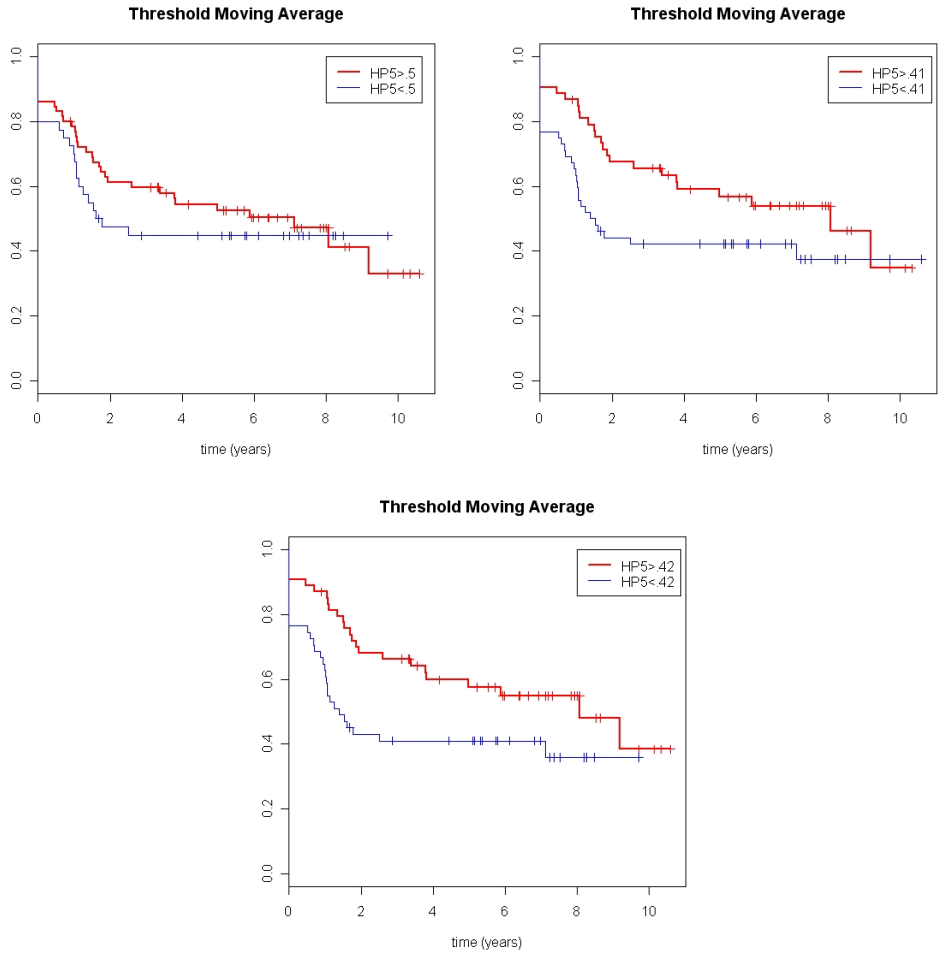


Figure F.6: Kaplan-Meier plots using the various different stratifications. The feature of note here is that the last plot, using .42 as the stratification, separates the two curves. Having curves that do not intersect is apparently beneficial because that means the Cox proportional hazards assumption holds [45][46].

	Log-rank P-value		
	Threshold choice for HP5		
	50%	Median	42%
Orginal Data	0.063	0.063	0.142
Moving Average	0.161	0.171	0.234
Threshold M.A.	0.473	0.062	0.030
Weighted M.A.	0.153	0.146	0.092

Figure F.7: P-values for the log-rank test using various stratifications and various filters. The results here seem to favour the threshold moving average filter and a stratification choice of .42, but we cannot conclude that such a filter and stratification threshold is useful in general because it is simply an ad hoc choice that works best for the data that I have.

F.3 Conclusion and Future Direction

Data collected by the Eppendorf needle is clearly noisy. There are massive oscillations between readings mere millimeters apart, physically impossible negative values of partial pressure that suggest at least a calibration error, and extremely high values of oxygen tension that suggest we may not be measuring inside the tumour. The first problem in particular motivates the use of a low-pass filter to try and eliminate the noise. The moving average filter is a simple filter that does a good job of smoothing the signal, makes sense physically, and is easy to implement, but the filtered data, for the most part, did not significantly alter the outcome of the statistical tests. We must be careful though, to not judge the merits of the filter based on the outcome of the log-rank test. We would certainly like the filter to lower the p-value and strengthen the result of the study, but in reality the data might not support such a conclusion and, as mentioned in the main text, a low p-value is not as important as many researchers believe. So to choose a filter that produced a lower p-value would be erroneous and misleading because in that case we would simply be manipulating data to create a certain result, and not analyzing the data to pursue the underlying truth.

The problem with using a filter is that we have nothing to guide us in our choice, there are literally hundreds of different filters that can be used. Even in the case of the moving average filter, different weights or thresholds or window sizes could potentially dramatically change the results. We could try to select filters based on what kind of noise we think the Eppendorf incurs, but given the inherent problems with measuring oxygen tension in tissue, that assumption might be hard to justify,

and may not narrow down the choices much.

Appendix G

Laplace's Rule of Succession

In the coin tossing example, we were able to estimate the bias parameter H which was our best estimate of the probability of getting a heads on any throw. However, if I were a gambler, what I would want to know specifically is the outcome of the next throw. More mathematically, one might ask:

> What is the probability that *the next throw* will be heads?

Of course, since we are Bayesians, a more accurate representation of the problem would be:

> Given that we have tossed a coin n times and gotten r heads. What is the probability that *the next throw* will be heads?

Laplace investigated a very similar problem many many years ago:

> Given that the sun has risen everyday for the past 5000 years, what is the probability that the sun will rise tomorrow?

Laplace's solution to the sunrise problem is known as "Laplace's Rule of Succession", and it must be pointed out here that the *only prior information* we are considering is *the number of times the sun has risen in the past 5000 years*. If we took into consideration other pieces of information, such as celestial mechanics, satellite data etc, the analysis would necessarily be altered, and it would be obvious that the sun will rise tomorrow, so the probability of the sun rising must be exactly (or extremely close to) 1, but that is precisely what we are not doing here, as we

wish to analyze a more general class of problems.

It occurred to me in the midst of investigating this problem that we might possibly use Laplace's result on the Eppendorf data. Recall previously that we estimated the HP5 of a patient exactly as though it were a Bernoulli trial, so Laplace's rule would necessarily apply to our data as well and predict the probability of whether the next measurement is going to be greater than or less than 5 mmHg. If we took a number of additional measurements and compared it to our prediction we might be able to verify whether our original estimate of HP5 was suitable or not.

(Historical note: Laplace postulated that the sun had risen for the past 5000 years because 5000 years was the commonly accepted age of the Earth at the time according to the Bible. We can adjust easily for personal experience, or for more recent estimates of the age of the Earth.)

G.1 Prediction vs Expectation

We want to assess what the probability will be that the next trial will yield a head given that we have already conducted n trials and gotten r heads. If the bias parameter H were known exactly, since we assume the trials are independent, this probability would clearly be exactly H. However, since H is unknown in our case and we are trying to estimate it, Jaynes [1] claims that the probability of interest is actually "numerically equivalent" to the expected value of the distribution we have derived for H, $< H >$. Jaynes proposes that this equivalence holds in very general settings, but does not seem to prove it. Instead, he provides us with a number of examples where the result holds.

G.1.1 Jaynes' illustrative example

Consider an urn with N balls, R of which are red. In the most general case, we do not know what the values of N and R are and wish to estimate them. To proceed, we would then draw n balls out of the urn, note that r of them are red, and analyze the data accordingly to get the joint distribution for N and R. However, let us suppose that we knew what N was and only wish to estimate R given our data. It will turn out that the result of interest, the predictive probability that the next ball drawn will be red, is independent of N, which is surprising, but very useful

because it means we do not need to know N after all.

Denoting the data that r balls out of n draws is red by D, and using Bayes' theorem, the probability distribution of R is given by

$$P(R|DNI) = P(R|NI)\frac{P(D|NRI)}{P(D|NI)}. \tag{G.1}$$

Suppose we were completely ignorant about R, then our prior must be the uniform distribution:

$$P(R|NI) = \begin{cases} \frac{1}{N+1} & \text{if } 0 \leq R \leq N, \\ 0 & \text{if } R \geq N. \end{cases} \tag{G.2}$$

(Note that we are in a discrete setting, hence the $N+1$ in the denominator.) For this urn problem, the likelihood function is the hypergeometric distribution [1]:

$$P(D|NRI) = \binom{N}{n}^{-1}\binom{R}{r}\binom{N-R}{n-r}. \tag{G.3}$$

The evidence, $P(D|NI)$, is a normalization constant as before, so we have

$$P(R|DNI) = S^{-1}\frac{1}{N+1}\binom{R}{r}\binom{N-R}{n-r}, \tag{G.4}$$

where S is the normalization constant. Jaynes [1] gives us this useful formula:

$$\sum_{R=0}^{N}\binom{R}{r}\binom{N-R}{n-r} = \binom{N+1}{n+1}. \tag{G.5}$$

Thus, we have

$$S = \sum_{R=0}^{N}\frac{1}{N+1}\binom{R}{r}\binom{N-R}{n-r} = \frac{1}{N+1}\binom{N+1}{n+1} = \frac{1}{n+1}\binom{N}{n}, \tag{G.6}$$

and hence

$$P(R|DNI) = \frac{n+1}{N+1}\frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}. \tag{G.7}$$

Now let's calculate the expected value of R as follows. We want to compute

$$< R >= \sum_{R=0}^{N} RP(R|DNI). \tag{G.8}$$

According to Jaynes, a useful formula is

$$(R+1)\binom{R}{r} = (r+1)\binom{R+1}{r+1}, \tag{G.9}$$

which yields

$$< R > +1 = (r+1)\binom{N+1}{n+1}^{-1}\binom{N+2}{n+2} = \frac{(N+2)(r+1)}{(n+2)}. \tag{G.10}$$

Thus the expected fraction of red balls left in the urn is

$$< F >= \frac{< R > -r}{N-n} = \frac{r+1}{n+2}. \tag{G.11}$$

Notice that this expectation is intuitively the probability that the next ball drawn will be red because it is the expected fraction of red balls to total balls remaining.

We shall derive the predictive probability that the next ball drawn will be red in a different manner using probability theory as follows. Let:

$$R_i \equiv \text{red ball on ith draw}, \quad 1 \leq i \leq N. \tag{G.12}$$

Then the probability of interest is, using marginalization,

$$\begin{aligned} P(R_{n+1}|DNI) &= \sum_{R=0}^{N} P(R_{n+1}R|DNI), \\ &= \sum_{R=0}^{N} P(R_{n+1}|RDNI)P(R|DNI), \\ &= \sum_{R=0}^{N} \frac{R-r}{N-n}\binom{N}{n}^{-1}\binom{R}{r}\binom{N-R}{n-r}, \\ &= \frac{r+1}{n+2}. \end{aligned} \tag{G.13}$$

Which yields exactly the same formula as the one for $< F >$. Jaynes notes that this equality is not a mere coincidence and that:

> "under quite general conditions, the *predictive probability* of an event at a single trial is numerically equal to the *expectation* of its frequency in some specified class of trials."

What exactly these conditions and trials are is not clear. Notice that the result is independent of N even though we assumed that we knew exactly what N was.

G.2 The Result

Going back to our coin toss problem, we want the probability that the next toss will be heads. According to the previous example, we would do so (that is, derive Laplace's rule of succession), by calculating the expected value of H. Recall from chapter seven that the posterior probability for the coin flipping problem has the following form (equation 6.9):

$$P(H|D,I) = \begin{cases} \frac{(n+1)!}{r!(n-r)!} H^r (1-H)^{n-r} & \text{if} \quad 0 \leq H \leq 1, \\ 0 \quad \text{otherwise.} \end{cases} \tag{G.14}$$

The expected value of H is then given by

$$\begin{aligned} P(H_{n+1}|DI) \; = \; < H > \; &= \; \int_0^1 HP(H|D,I)dH, \\ &= \; \int_0^1 \frac{(n+1)!}{r!(n-r)!} H^{r+1}(1-H)^{n-r} dH, \\ &= \; \frac{r+1}{n+2}. \end{aligned} \tag{G.15}$$

The integral was evaluated using the known result for beta functions (equation 6.7) as in chapter 6. The result for the coin flip is identical to the previous one (G.13) because it is simply the continuous version of the discrete result as $N \to \infty$. Since the formula is independent of N, this limit has no effect on the outcome.

G.2.1 So will the sun rise tomorrow?

Assuming the sun has risen everyday for the last 5000 years, we have that $n = r = 5000 \times 365$, so

$$P = \frac{5000 \times 365 + 1}{5000 \times 365 + 2} \approx 0.9999995. \tag{G.16}$$

This means that even the most naive estimate would have the sun almost certain to rise tomorrow, though, of course, aliens might invade and destroy the sun before then!

G.2.2 Back to Eppendorf

My original idea was the following: Suppose it were possible to get additional Eppendorf measurements from the patients. Then we can test the accuracy of the

Study Number	HP5			LRS	Median PO2	MaxEnt Median
1	0.520	±	0.050	0.520	4.7	14.1
2	0.438	±	0.044	0.438	12.45	28.5
3	0.568	±	0.081	0.564	3.2	8.5
4	0.468	±	0.063	0.469	6.75	10.0
6	0.917	±	0.024	0.911	0.6	2.9
7	0.127	±	0.033	0.135	31.85	51.8
8	0.325	±	0.052	0.329	28.75	50.7
10	0.825	±	0.035	0.820	0.9	2.2
12	0.578	±	0.052	0.576	2.45	7.6
13	0.706	±	0.038	0.703	3.4	7.1
14	0.172	±	0.033	0.176	56.35	68.5
15	0.123	±	0.043	0.136	66.9	76.1
17	0.357	±	0.043	0.359	10.4	20.4
18	0.649	±	0.039	0.647	1.5	2.7
19	0.589	±	0.052	0.587	3.2	4.5
20	0.549	±	0.043	0.548	3.5	4.2
21	0.126	±	0.029	0.132	30.7	53.0
22	0.302	±	0.070	0.311	16.7	36.1
23	0.470	±	0.041	0.471	5.5	14.8
24	0.635	±	0.045	0.632	2.3	3.0
25	0.874	±	0.032	0.867	1.4	4.9
27	0.260	±	0.040	0.264	32	11.3
29	0.630	±	0.040	0.628	3	4.9

Figure G.1: Here is a comparison between the results for Laplace's rule of succession (LRS) and the best estimate for HP5 from chapter 6. For completeness, the sample median and maxent median estimates are also included.

HP5 estimate by taking a number of additional readings, and comparing the probability of getting a reading less than 5mmHg in these new readings compared to the predicted probability of getting a reading less than 5mmHg using Laplace's rule of succession.

However, a quick glance at the table of results for Laplace's rule of succession compared to the previous results from chapter 6 for the best estimate (figure G.1) will show that the two are nearly identical for all patients. At the very least, the Laplace result is within the error bar for the best estimate. The reason for this is mostly likely because of the large number of data that we had, which meant that there was little difference between the best estimate, given by r/n, and Laplace's rule of succession, given by $r+1/n+2$. So while the sunrise problem is an interesting intellectual puzzle, it does not yield particularly useful information for our data.

References

[1] E. T. Jaynes, *Probability Theory - The Logic of Science*. Cambridge University Press, Cambridge, 2007. 1, 64, 66, 122, 124, 141, 142

[2] D. S. Sivia. *Data Analysis - A Bayesian Tutorial 2nd Ed.*. Oxford University Press, New York, 2006. 1, 2, 5, 66, 68, 79, 117, 119, 122

[3] G. Larry Bretthorst, *Bayesian Spectrum Analysis and parameter Estimation.* Lecutre Notes in Statistics. Springer-Verlag, New York. 1988. Provided online by Bretthorst at http://bayes.wustl.edu/ 1

[4] Donald A. Berry, *Bayesian Clinical Trials*. Nature Reviews, vol. 5, 2006. 1

[5] *Can Bayesian Approaches to Studying New Treatments Improve Regulatory Decision Making?* U.S. FDA workshop held May 20-21, 2004 at Johns Hopkins University. Presentations available at http://www.cfsan.fda.gov/~frf/bayesdl.html 1

[6] *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials - Draft Guidance for Industry and FDA Staff.* Draft released for comment on May 23, 2006. http://www.fda.gov/cdrh/osb/guidance/1601.html 1

[7] Paul Graham, *Better bayesian Filtering.* Jan 2003, http://www.paulgraham.com/better.html 63

[8] Stephen F. Gull, John Fielden. *Bayesian Non-Parametric Statistics* 88, 89

[9] Eric J. Hall, *Radiobiology for the Radiologist. 4th ed.* Lippincott Company, Philadelphia, 1994. 9, 10, 12, 13, 14, 15

[10] L. Harrison, M. Chadha et al, *Impact of Tumour Hypoxia and Anemia on Radiation Therapy Ouctomes.* The Oncologist, 2002, 7:492-508. 9, 15

[11] A. Fyles, M. Milosevic, et al, *Tumor Hypoxia has Independent Predictor Impact Only in Patients With Node-Negative Cervix Cancer.* Journal of Clinical Oncology, 2002, vol.20, no.3, 680-687. 6, 9, 15, 33, 35, 36, 37, 39, 42, 44, 48, 49, 56, 57, 58, 82, 87

[12] R. Wong, A Fyles, et al, *Heterogeneity of Polarographic Oxygen tension Measurements in Cervix Cancer: An Evaluation of Within and Betwenn Tumor Variability, Probe Position, and Track Depth.* Int. J. Radiation Oncology Biol Phys., 1997, vol.39, no.2, 405-412. 7, 9, 28, 33, 35, 36, 37, 38, 49

[13] A. Fyles, M. Milosevic, et al, *Long-term performance of interstial fluid pressure and hypoxia as prognostic factors in cervix cancer.* Radiotherapy and Oncology, 2006, vol.80, 132-137. 59

[14] P. Stadler et al, *Influence of The Hypoxic Subvolume On The Survival of Patients with Head And Neck Cancer.* Int. J. Radiation Oncology Biol Phys., 1999, vol.44, no.4, 749-754. 7, 9, 36, 49

[15] P. Vaupel et al, *Oxygenation of Human Tumors: Evaluation of Tissue Oxygen Distribution in Breast Cancers by Computerized O_2 Tension Measurements.* Cancer Research, 51, 3316-3322, June 15, 1991. 7, 9, 35, 36, 37, 49

[16] G. Stüben et al, *Experimental Studies on the Possible Influence of Invasive Oxygen Measurements on Tumour Radiosensitivity.* Acta Oncologica, 1998, vol. 37, no. 4. 369-373. 7, 9, 38, 39, 49

[17] T. H. Knocke, H. D. WEitmann, et al., *Intratumoral PO2 measurements as predictive assay in the treatment of carcinoma of the uterine cervix.* Radiotherapy and Oncology, vol. 53, 1999, 99-104. 7, 9, 44, 49

[18] D. R. Collingride, J. M. Piepmeier, et al. *Polarographic measurements of oxygen tension in human glioma and surrounding peritumoural brain tissue.* Radiotherapy and Oncology, vol. 53, 1999, 127-131. 7, 9, 44, 49

[19] M. Nordsmark, J. Loncaster, et al., *Measurements of hypoxia using pimonidazole and polarographic oxygen-sensitive electrodes in human cervix carcinomas.* Radiotherapy and Oncology, vol. 67, 2003, 35-44. 16, 44, 49

[20] I. Toma-Daşu et al, *Conversion of polarographic electrode measurements-a computer based approach.* Phys. Med. Biol., 2005, vol. 50, 4581-4591. 39, 40

[21] I. Toma-Daşu et al, *Theoretical simulation of oxygenn tension measurement in tissues using a microelectrode: I. The response function of the electrode.* Physiol. Meas., 2001, vol. 22, 713-725. 39, 40

[22] J. Scott Armstrong, *Significance tests harm progress in forecasting.* International Journal of Forecasting, 2007, vol.23, 321-327. 59, 60

[23] J. Scott Armstrong, *Statistical Significance Tests are Unnecessary Even When Properly Done.* International Journal of Forecasting, 2007, vol.23, 335-336. 59, 60

[24] I. Fatt, *Polarographic Oxygen Sensors.* CRC Press 1976. 17, 18, 19, 20, 22, 23, 27, 28, 32, 37, 38, 107

[25] P. W. Davies, *The Oxygen Cathode.* Physical Techniques in Biological Research, vol. IV, ed. W. Nastuk. Academic Press 1961. 137-179. 23

[26] I. M. Koltholf and J. J. Ligane, *Polarography - Volume I, Theoretical Principles; Instrumentation and Technique.* Interscience Publishers, New York, 1952. 19, 23, 28, 32

[27] K. H. Mancy et al, *A Galvanic Cell Oxygen Analyzer.* J. Electoanal. Chem., 4, 1962, 65-92. 21, 32

[28] R. A. Albanese, *Use of Membrane-covered Oxygen Cathodes in Tissue.* J. thoer. Biol., 1971, vol.33, 91-103. 22, 23, 24, 28, 30, 32, 37, 104, 107

[29] C. E. W. Hahn, *Electrochemical analysis of clinical blood-gases, gases, and vapours.* Analyst, June 1998, vol. 123, 57R-86R. 19, 28, 33, 34, 36, 37, 38, 39, 42

[30] J. A. Hudson, *Measurement Of Oxygen Tension By The Oxygen Cathode.* Med. & biol. eng., 1967, vol. 5, 207-223. 22, 38, 107

[31] I. A. Silver, *Polargraphy and its biological applications.* Phys. Med. Biol., 1967, vol.12, no.3, 285-299. 28, 32, 33, 34, 35, 38

[32] D. L. Short and G. S. G. Shell, *Fundamentals of Clark membrane configuration oxygen sensors; some confusion clarified.* CSIRO Centre for Irrigation Research, Australia, Feb, 1984. 32

[33] H. S. Carslaw, J. C. Jaeger, *Conduction of Heat in Solids.* 2nd Ed. Oxford University Press, 1959. 25

[34] J. J. Carroll, *What Is Henry's Law?* Chemical Engineering Progress. Vol. 87, no.9, 1991, 48-52. 107, 110, 111

[35] J. J. Carroll, *Henry's Law Revisited.* Chemical Engineering Progress. Jan. 1999, 49-56. 109, 111

[36] F. L. Smith, A. H. Harvey, *Avoid Common Pitfalls When Using Henry's Law.* Environmental Management. Sep 2007, 33-39. 111

[37] Univeristy of Delaware Physical Chemistry lecture notes. http://www.udel.edu/pchem/C443/Lectures/Lecture33.pdf 109

[38] Universtiy of Florida Physical Chemistry lecture notes. http://www.chem.ufl.edu/~itl/4411/colligative/lec_i.html 108

[39] Univeristy of Arizona Chemistry 480A notes. Aug 2001. http://www.chem.arizona.edu/~salzmanr/480a/480ants/chemther.html 108, 112

[40] Steven W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing.* http://www.dspguide.com 131

[41] Rakesh K. Jain, *Normalizing tumor vasculature with anti-angiogenic therapy: A new paradigm for combination therapy.* Nature Medicine, vol. 7, no.9, Sep 2001, 987-989. 10, 11

[42] T. Roose, J. Chapman, et al, *Mathematical Models of Avascular Tumor Growth.* SIAM Review, vol. 49, no.2, 179-208 10, 11

[43] *Answers to Your Radiation Therapy Questions.* American Society of Radiation Oncology (ASTRO) http://www.rtanswers.org/ 9

[44] R. Peto, M. C. Pike, et al, *Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient.* Br. J. Cancer, 1977, 35, 1-38. 57, 115

[45] D. Collet, *Modelling Survival Data in Medical Research. 2nd Ed.* Chapman & Hall/CRC, USA, 2003. 50, 51, 52, 54, 55, 56, 115, 116, 137

[46] M. Pintilie, *Competing Risks - A Practical Persepctive.* Wiley. Ontario Cancer Institute, Canada. 50, 54, 55, 56, 115, 137

[47] J.H. McDonald, *Handbook of Biological Statistics.* Sparky House Publishing, Baltimore, Maryland, 2008. http://udel.edu/~mcdonald 48

[48] E. L. Kalplan and P. Meier, *Nonparametric Estimation From Incomplete Observations.* American Statistical Association Journal, June 1958, 458-481. 52

[49] J.D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference, 3rd Ed.* Marcell Dekker, 1992. 113, 114

[50] Frank E. Beichelt and L. Paul Fatti, *Stochastic Processes and Their Applications.* Tayloer & Francis, New York, 2002. 94, 95

[51] Blood Gases, Medline Plus, U.S. National Library of Medicine http://www.nlm.nih.gov/medlineplus/ency/article/003855.htm 42

[52] Bob Bartlett, *Answering questions about hypoxia.* Wound Healing Persepctives, vol. 1, no.4, 2004. 43, 44