The London School of Economics and Political Science

The Problem of Model Selection and Scientific Realism

Stanislav Larski

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, UK, February 2012

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others [in which case the extent of any work carried out jointly by me and any other person is clearly identified in it].

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 40743 words.

Abstract

This thesis has two goals. Firstly, we consider the problem of model selection for the purposes of prediction. In modern science predictive mathematical models are ubiquitous and can be found in such diverse fields as weather forecasting, economics, ecology, mathematical psychology, sociology, etc. It is often the case that for a given domain of inquiry there are several plausible models, and the issue then is how to discriminate between them – this is the problem of model selection. We consider approaches to model selection that are used in classical [also known as frequentist] statistics, and fashionable in recent years methods of Akaike Information Criterion [AIC] and Bayes Information Criterion [BIC], the latter being a part of a broader Bayesian approach. We show the connection between AIC and BIC, and provide comparison of performance of these methods.

Secondly, we consider some philosophical arguments that arise within the setting of the model selection approaches investigated in the first part. These arguments aim to provide counterexamples to the epistemic thesis of scientific realism, viz., that predictively successful scientific theories are approximately true, and to the idea that truth and predictive accuracy go together.

We argue for the following claims: 1) that none of the criticisms brought forward in the philosophical literature against the AIC methodology are devastating, and AIC remains a viable method of model selection; 2) that the BIC methodology likewise survives the numerous criticisms; 3) that the counterexamples to scientific realism that ostensibly arise within the framework of model selection are flawed; 4) that in general the model selection methods discussed in this thesis are neutral with regards to the issue of scientific realism; 5) that a plurality of methodologies should be applied to the problem of model selection with full awareness of the foundational issues that each of these methodologies has.

Acknowledgements

To the loving memory of my grandparents

I owe an enormous debt of gratitude to my supervisors Colin Howson and Roman Frigg. It is a brute fact that without their encouragement, wholehearted support and uncountably infinite patience this thesis would not have seen the light of day.

I feel greatly privileged to have had all of my university education at the Department of Philosophy, Logic and Scientific Method in particular and at the London School of Economics and Political Science in general. It is truly my intellectual *Alma Mater*.

I am grateful to my friends and fellow PhD students Arhat Virdi, Bengt Autzen, Fernando Morett, Gary Jones, Hamid Sayamdost, Ittay Nissan-Rozen, Jeremy Howick, Kizito Kiyimba, Lefteris Farmakis, Matteo Morganti, Michal Polak, Michael Stentenbach, Sheldon Steed, most of whom dwelled to various degrees in the Philosophy Basement. I shall miss the unique atmosphere of intellectual curiosity, mutual support and sheer fun.

Last but by no means least my huge thanks go to my family who have on par with me undertaken this arduous yet fulfilling journey.

Contents

1. Introduction and Classical Methods of Model Selection

1.1 The Three Problems of Model Construction

In life in general, and in science in particular, one is often interested in issues as to how to explain, or predict various phenomena. For instance, where would the cannon ball fall if one were to shoot it from a certain cannon? And, for that matter, what is the explanation as for why it is to fall [or has already fallen] in the predicted place [or, indeed, elsewhere]? Explanation is a fascinating subject in itself. However, in this thesis we shall concentrate on the no less fascinating subject of scientific [more specifically, statistical] prediction.

So, let us get to the cannon ball example. How are we to predict where the cannon ball is to land if shot? In order to do so we come up with a model. That is, we engage into the process of abstraction and idealisation from the 'real world'. We abstract from the features of the world that are deemed irrelevant for our purposes and take into account only the relevant facts according to Newton's physics [which we will take for granted in this example] such as the angle of elevation of the cannon with respect to the ground level, the velocity of the cannon ball as it exits the barrel of the cannon, the weight of the cannon ball, the speed and direction of wind, friction in the barrel of the cannon, etc. We idealise certain features. For example, it may be impractical and costly to calculate the friction within the cannon's barrel as it is, so we may assume that it is a totally smooth surface. We make further assumptions such as that the speed and direction of wind are both constant. We may sketch our model on the back of the envelope for ease of representation. Once we have done all of these, we have our predictive model.

For our purposes we can think of a scientific model as a tool, which aids us in generating predictions of phenomena of interest. We are certain that the cannon ball model [quite possibly not in exactly the same way as it is envisaged here, but, we would venture, closely enough] was an important predictive tool utilised by the Western armies of a couple of centuries back.

At this junction let us draw an important distinction between theoretical and statistical modelling. This is not the only distinction one can draw, and for the purposes of this thesis we shall use it rather loosely, for we are concerned with statistical modelling that has theoretical influence/elements in it. However, this distinction provides conceptual clarity to our proceedings. A theoretical model is a model constructed using a general theory without involving data. The preceding example of a model is in fact an example of theoretical modelling. In this example such theory is Newtonian mechanics. We feed the initial conditions into the equations of Newton's mechanics to yield our prediction. However, this thesis is going to be concerned with statistical models. These models are predominantly built from the data upwards without much use of the general theory, if any. To illustrate, let us use the setup of the cannon ball example. If we wanted to construct a 'pure' statistical predictive model, we would shoot several cannon balls from the cannon every time observing the quantities that we consider relevant such as the amount of gun power input, the angle of elevation of the cannon barrel, exit velocity, velocity and direction of wind, etc. When shooting cannon balls we would vary the relevant quantities – e.g., we would vary the amount of gunpowder, change the angle of elevation, etc. to see how it affects the distance that our cannon balls travel. Then we would come up with a model by means of correlating these data. We imagine [although we have not undertaken research into this matter] that early Chinese users of cannon technology and the medieval Western armies would have modelled the phenomenon in a way akin to our description of statistical modelling[1].

There is also a salient distinction within modelling between deterministic and probabilistic models. It has to be emphasised that this distinction is independent of the theoretical vs. statistical distinction. Deterministic models are such that the predictions that they issue are of a definitive nature. For example, a deterministic model may predict that given the current amount of gunpowder, the elevation of the cannon barrel and the velocity and direction of wind, the cannon ball will land exactly 552 metres due north if shot now[2]. Whereas, using the same example, the

[1] Reader interested in the historical development of projectile technology is referred to Crosby (2002).
[2] For our purposes we take Popper's definition of scientific determinism, viz.: '…the doctrine that the state of any closed physical system at any given future instant of time can be predicted, even from within the system, with any specified degree of precision, by deducing the prediction from theories, in conjunction with initial conditions whose required degree of precision can always be calculated [in

probabilistic variant thereof would yield a distribution of likely landings with probabilities attached to them. The following table provides examples of each of the four types of models.

Models	Theoretical	Statistical
Deterministic	Cannon model constructed by using general theory which issues definitive "non-chancy" predictions	Cannon model constructed by correlating data which yield a definitive curve such that all the data points lie on it
Probabilistic	A model of radioactive decay of radioactive elements.	Cannon model constructed by correlating data which yield a definitive curve such that the data points lie close to it reflecting imprecision of measurement

It is clear that the cannon ball model fits into the deterministic theoretical category. On the other hand, an example of a theoretical probabilistic model is the radioactive decay model, which is solidly based on the theory of quantum mechanics that issues probabilistic predictions. Deterministic statistical models, although logically possible, are in practice rather fictitious, for their construction involves highly restrictive conditions. For instance, in our cannon ball example, the cannon would have to be fired indoors to remove the factor of the wind, or a deterministic theory of the wind movement would have to be added, which on current scientific thinking is not feasible, because, among other conditions, it requires infinite precision of measurement of the initial conditions[3]. Still, we shall use deterministic statistical models for ease of introduction to the issue of model selection among probabilistic statistical models.

accordance with the principle of accountability] if the prediction task is given.' [Popper (1982):36] For a thorough discussion of determinism cf. Earman (1986).

[3] For a thorough introductory text on the mathematical chaos theory, of which this is an instance, please see Stewart (2002).

Indeed, in this thesis we concentrate on statistical probabilistic models. The reason for the focus on statistical modelling is that such models have gained prominence and play an enormous role in many sciences. The list of sciences that use statistical modelling keeps growing. It finds application in economics, sociology, mathematical psychology, environmental sciences, etc.

A probabilistic statistical model is a mathematical equation, with the aid of which one describes the phenomenon under study in terms of random variables that have probability distributions ascribed to them. The explanation of what these terms are will be provided in section 1.3. As we mentioned at the very beginning of this chapter, in this thesis we concentrate on use statistical models for the purposes of prediction.

Let us now turn to statistical models and see how they are constructed using a simple example.

Suppose, for instance, that we are interested in finding out how the heights and weights are correlated with each other of, say, males, who are in their 20's and who live in the London borough of Waltham Forest. The reason for such a fascination with the heights and weights could be that we are perhaps acting on behalf of the local health authority, which is in the process of planning a new hospital. The authority may be interested in obesity [e.g., they may want to predict what the Body Mass Index[4] within the Borough would be], or in predicting as to what would be the optimal height of the doorways, the sizes of beds, weight load of equipment such as wheel chairs. They may also hold a general interest in the demography of the Borough.

Let us suppose that we would like to predict the weight of any such male given his height. In order to draw an inference we need to do three things. First, we collect a sample of data from the population. Second, we choose the structure of the model

[4] BMI is one of the most widely recognised indices used in order to classify weight of adults. It is defined as weight (kilograms) / height^2 (metres). If one's BMI is below 18.5, one is considered to be underweight (in particular, if BMI < 16, one is classified as "severely thin") whereas if one's BMI > 25 one is considered to be overweight (in particular, if BMI > 30, one is classified as "obese"). Source: World Health Organization: http://apps.who.int/bmi/index.jsp?introPage=intro_3.html

[that is, the functional form of the model or, in other words, the family of models which have the same functional form but differ in that their parameters are set at different values] according to which the weights and heights are related. Third, having chosen the structure, we determine the values of parameters, that is, we pick a particular model[5] from the family of models. Let us consider these steps in turn.

1.1.1 Sampling

This section is here solely for completeness of presentation of statistical modelling process. The focus of the thesis shall be entirely on the issue on model selection and on parameter estimation. We will be concerned with parameter estimation insofar as it is relevant to model selection. Hence we gloss over quite interesting issues in sampling[6]. We mention the solutions that we find reasonable and appealing without much argumentation in order to give the reader a sense of where we stand on these issues.

The question as to how to draw such a sample properly has attracted a lot of attention in statistics. Sampling techniques can be divided into two categories – random sampling and judgement [representative] sampling[7].

1.1.1.1 Random Sampling

In random sampling every member of the population has to have a known objective probability of being selected for sampling to be called random. In our example one way that this can be achieved is by assigning every known male in the borough of Waltham Forest a unique natural number, then putting each number on a separate ball, then placing all the balls in an urn and drawing *n* balls [*n* corresponding to the size of the sample] from the urn without looking [so that each ball has an equal

[5] Our usage of the term 'model' here closely follows van Fraassen's: 'Thus in the scientists' use, 'model' denotes what I would call a model-type. Whenever certain parameters are left unspecified in the description of a structure, it would be more accurate to say … that we described a structure-type. I will continue to use the term 'model' to refer to specific structures, in which all relevant parameters have specific values.' [van Fraassen (1980):44] Our notion of a model corresponds to van Fraassen's 'model-type' or 'structure-type'.
[6] For further details and discussion of sampling cf. Stuart (1962), Stuart (1984), Urbach (1989).
[7] *Ibid.*

chance of being picked], noting down the numbers and contacting the individuals who had those numbers associated with them to find out what their heights and weights are. In fact, this is an example of simple random sampling, where every member of the population has an equal probability of being selected.

A different way to do random sampling would be to divide the population into sub-populations [strata] with respect to some characteristics that are believed to be correlated with the attributes of primary interest. So, in our weights/heights example, weights and heights of individuals are such attributes of primary interest, and the characteristics according to which the population of males can be divided could be the countries of their origin [for instance, it is commonly observed that males from Scandinavian countries tend to be relatively tall and slender, and, say, males from the Indian subcontinent also tend to be slender, but are relatively shorter than the Scandinavians], the level of their disposable income [males on the relatively lower incomes seem to consume more unhealthy foods], etc. Once the population is stratified in this way, the simple random sampling is done within each stratum. The merit of stratified sampling in comparison to simple random sampling is that in situations where there is at least some amount of prior knowledge about possibly correlated characteristics, stratification results in more precise estimation [i.e., inferences from stratified samples almost always have smaller variance – the measures of precision are to be discussed in subsequent sections]. Stratification maximises precision when the average values of observations are as different as possible, and their variances are as small as possible[8]. Intuitively, the maximal difference implies that the characteristic according to which the stratification was done is correlated with the attributes of interest. In fact stratification with respect to any characteristic leads to an increase in precision, so long as the size of the sample is small in proportion to the population, and the strata contain more than one member.

Another type of random sampling is cluster sampling. In cluster sampling one also divides the population into sub-populations, but instead of doing random sampling within each sub-population, one randomly selects a single sub-population, and then

[8] cf. Stuart (1962):49

makes up the sample from all the individuals within the selected sub-population. An example of cluster sampling is list sampling. If we take the list of all the relevant males in our particular example in alphabetical order of their surnames, then divide the population with respect to the first letter of their surname in such a way that the number of individuals in each cluster is about the same [so if the number of individuals that have the letter S as the first letter of their surname is about the same as the number of males that have their surname begin with X or Y or Z, then we form two clusters – one S cluster and one XYZ cluster, and carry on in this fashion with respect to the other letters of the alphabet] and then randomly select one such sub-population to constitute our sample, then we will have done cluster sampling. An advantage of cluster sampling over stratified sampling is that sometimes population is naturally arranged into clusters – for example, into districts, or households, into groups of employees or different companies, etc. On the other hand, for cluster sampling to achieve an improvement in precision over stratified sampling, the individuals within the clusters have to be maximally varied. Intuitively that means that clusters should be as representative of variation within the population as possible. So, following the earlier example, if we are to do list sampling, under each first letter of a surname we would like to have some Scandinavians, some males from Indian subcontinent, etc., in our clusters roughly in proportion in which they occur in the whole population. If, however, our clusters are not varied, cluster sampling achieves much lower precision than both simple random and, a fortiori, stratified sampling. That is, if, say, the cluster ABC is randomly selected, and it so happens that young adult males from Scandinavia predominantly have such surnames, then we would have a sample skewed towards relatively slim tall males.

1.1.1.2 Judgement Sampling

Judgement [also known as representative] sampling is the same as stratified random sampling, but for one important feature – it is not random. The idea behind judgement sampling is that the most important thing that one [that is, a researcher who does sampling] has to do is to choose according to which categories the population should be divided into sub-populations. Once that is done, one then determines how many individuals should be ‘observed’ in each sub-population based on the proportion of the quantity of individuals in a given sub-population with

respect to the total number of individuals in the population. Then one picks the determined number of individuals in each sub-population [hence it is sometimes referred to as quota sampling] in whatever way it is most practicable to do so – randomisation in this case is not the *sine qua non*.

Let us further clarify what the difference between stratified and judgement sampling is. Indeed, it is the case that in both methods one divides the population into sub-population according to some salient characteristics. However, in stratified random sampling one has to draw samples from sub-population by randomised sampling, whereas in judgement sampling one is free to pick individuals for one's sample according to one's own ideas.

1.1.2 Model Selection

So, suppose that we have picked a sample in one of the ways described in the sub-section above. What do we need to do further? We need to choose [or construct] a statistical model, which involves choosing the mathematical structure, and pick the values of parameters. In our usage 'model selection' refers to choosing the mathematical structure. The issues of what scientific models are, how they interact with theories and observations, etc. have attracted a lot of attention in the recent years[9]. However, as we mentioned in the beginning of this section, in this thesis we will consider statistical models only.

Now, why do we need a statistical model in our example? Since we are interested in finding out the relationship between weights and heights of the males [say, we are trying to come up with a generalisation for the purposes of prediction as to what the height of any such male within the Borough will be, given his weight], we would like to know the form of this relationship. That is, for a given unit increase in height of a male, would his weight be expected to increase in linear proportion, or perhaps quadratic, or cubic, or in some other way? Would a unit change in height correspond to the same change in the weight if the person is relatively 'tall' rather than if he is

[9] For a comprehensive survey cf. Frigg and Hartmann (2006).

somewhat 'short'? To begin with, let us see what form statistical models can take in order to make sense of the model selection approaches.

For now we will introduce deterministic statistical models, since they are in a sense simpler than probabilistic statistical models. It shall be easier to move onto probabilistic statistical models once we consider deterministic ones because these two types of models have many features in common.

$Y = aX + b$ is an example of a linear model [call it LIN]. Each combination of the values of parameters a and b would pick out a particular element within the linear model – an element of LIN. LIN has two variables – X is usually referred to as the independent variable and Y as the dependent variable. To make this model probabilistic one would need to introduce a random component [it is also often called an error term] ε: $Y = aX + b + \varepsilon$, where ε has a probability distribution[10]. Another example of a deterministic statistical model is the quadratic one [call it PAR]: $Y = aX^2 + bX + c$. The elements of PAR for which $a \neq 0$ are represented by parabolic curves in the Cartesian plane. Since in our example we are interested in predicting the weight, the dependent variable Y represents the weight measured, say, in kilograms, and the independent variable X represents the height measured, say, in centimetres.

Now, the two schools of statistical thought within which the vast majority of statistical reasoning takes place are the so-called Classical statistics and Bayesian statistics. We defer consideration of Bayesian statistics until chapter 4.

In chapter 2 we consider some of the methods of Classical statistics. These methods are not traditionally thought to be about model selection, although they can be viewed as such, at least to a limited extent [cf. Forster (2001)]. Roughly speaking, the methods of Classical statistics usually assume that the functional form of a hypothesis [or, in our usage above, a model] is known, and proceed to use samples of data to test models with the parameters set at particular values either by themselves or against an alternative model with different values of parameters, or to

[10] This notion, among others, will be elucidated in section 1.3.

test two subsets of the same model against each another, or estimate parameters from samples of data by particular values [thus picking out an element of the model] or by ranges of values [thus narrowing the range of plausible elements within the model].

The reasons as to why we consider Classical statistical methods even though they are related to model selection in a rather limited sense are the following. Firstly, Classical statistics is the most influential type of statistical reasoning, familiarity with at least the major points of which is pre-requisite for any field of statistical analysis. Secondly, the methods of Classical statistics are used by many as the 'gold standard' against which all other methods are judged, including the methods which we consider in chapters 3 and 4, that take model selection as their explicit aim. Thirdly, the methods of Classical statistics have featured in the philosophical debate with regards to the putative connection between model selection methods and scientific realism, to the consideration of which we turn in chapter 5.

In section 4.1 we consider the main features of Bayesian statistics, which has been the main rival to the Classical statistical thought in modern statistics[11]. In Bayesian statistics the issue of model selection arises quite naturally.

1.1.3 Parameter Estimation

At this point let us state that throughout this thesis we are concerned with *parametric* modelling. That is, with models which have finite-dimensional vector-valued parameters. For non-parametric methods see Silvey (1975):chapter 9 and Spanos (2001).

As we mentioned in section 1.1.2, choosing a statistical model amounts to choosing a set of mathematical equations that have the same structure. E.g., $Y = aX + b$ is a linear model specifying an uncountably infinite set of particular lines that have distinct values of parameters a, b. As we noted above, this linear equation does not amount to a *probabilistic* statistical model [it lacks a random component as it stands] but that will matter later on in the thesis. For the ease of introduction a deterministic

[11] For an insightful summary of the debates both internal and external to the Classical statistics see Mayo (2005).

statistical model will do. Once we have picked/found our statistical model [suppose for now that we picked the linear model in our height/weight example, where X denotes the heights variable and Y denotes the weights variable], the task is to estimate the values of the parameters *a* and *b* from the sample data that we have. These values would give us a particular statistical model [that is, a particular element of linear model]. A formula whereby estimation is carried out is called an estimator, whereas the particular values that it takes are called the estimates. Logically there are infinitely many ways of doing so. Let us briefly see how the Classical and Bayesian approaches attempt to solve the issue. We shall go deeper into the Bayesian approach in section 4.1. The introduction below is conducted in very general terms because the definition and explanation of statistical terms necessary for more precise rendition is forthcoming in later sections.

Classical statistics has a list of properties that an admissible estimator should have. The most important and most commonly used properties are unbiasedness, consistency, efficiency and possession of minimum squared error. Let us look at these in turn.

An estimator is unbiased when the estimates that it yields across different samples are on average equal to the value of the true parameter. An estimator is said to be consistent when, as the sample size tends towards infinity, the estimates provided by the estimator converge on the true value. An estimator is efficient just in case the estimates yielded from the estimator have the minimum spread among the estimators within the same class. That is, the range within which such estimates lie is on average the shortest [in statistical terminology, this is expressed as the estimator has the minimum variance]. Here is an example. Let us go back to the linear model $Y = aX + b$. Let us suppose that we want to estimate the value of the parameter *a*. In classical statistics we assume that the value of *a* is fixed but unknown. How should we go about the estimation? Again, without getting into the formal details, one way to do so would be this. We can plot the data points in the Cartesian plane and draw a line [that would be a particular manifestation of the linear model] in such a way that the sum of the squared vertical distances [that is, along the y-axis] from each point to the line is minimised. Thus this line would lie closer to each data point than any other element of the linear model [in the sense of minimal vertical square distance].

The reasoning behind adopting such a method is that presumably our model should reflect the data as closely as possible in order to have any predictive success. We shall return to this point in chapter 3.

Suppose now that our line is $y = \alpha x + \beta$, where α and β are such that the line $Y = \alpha X + \beta$ has the minimal sum of square distances to all data points within the sample. Now, suppose that we are restricting our attention to the group of linear estimators. That is, we are to pick estimators of a among the functions $\hat{a} = c\alpha + d$, ($\hat{a}$ stands for an estimator of a), so that a is a linear function of α. Now, what would be the best linear estimator among the infinitely many? The "classical" answer is that the best one is where c = 1 and d = 0. That is, $\hat{a} = \alpha$. It is demonstrated that this estimator is unbiased, consistent and, under further conditions known as Gauss-Markov conditions [which there is no need to go into at this point], it has the minimum variance, i.e., that it is efficient.

In Bayesian statistics point estimates are generally not provided because the inference is based on the full posterior distribution[12], but point estimates can be derived. One popular method is called MAP – maximum a posteriori. Under this method the point estimator of a parameter is such that it provides the maximum posterior probability of the model in the light of the sample. This is equal to the mode of the posterior distribution. The mode of any sample is the value of random variable that occurs most frequently. To give a simple example, suppose that we rolled a die 7 times, and that the following is our sample of numerical outcomes: {1, 1, 2, 4, 5, 5, 5}. In this case the mode is 5.[13]

There is also the method of Maximum Likelihood Estimation. We defer consideration of this method until chapter 3, because understanding it will be crucial for the discussion of the Akaike Information Criterion in that chapter.

[12] Roughly speaking, posterior probability distribution comprises a set of probabilities associated with each possible value of the parameters within a model in the light of data. We say more on this point in section 4.1.

[13] If the distribution is symmetrical univariate [i.e., it has only one random variable in it; we will see that the normal distribution is an example of such a distribution], the mean, mode and median are the same. The median of any sample is the middle value when the values arranged from the smallest to the largest in order. In this case the median is 4.

1.2 Methodological Issues

1.2.1 Sampling

Please note that this section is here solely for the purpose of completeness of introducing the issue of model selection. The issues with sampling will not be considered in the rest of the thesis. It will be assumed that our data were gathered by some satisfactory method. So the issues in this section are flagged for possible interest of the reader, and some signposts are indicated as to where our philosophical opinion lies without much argument for or against, which is done deliberately.

> '…Principle of Random Sampling asserts that satisfactory estimates can only be obtained from samples that are objectively random…'
>
> Howson and Urbach (2006):178

The primary motivation for random selection of individuals to constitute the sample is that such a selection allows one to obtain a sample free of biases. A salient example of a possible bias is the selection bias. That is, conscious or unconscious tendency on behalf of the researcher to select members for the sample on the basis of some subjective idea as to what the salient characteristics of the population are. In random sampling what is important is the procedure whereby the sample is chosen, and not the actual outcome. The procedure has to be fair. That is, paradoxically [and it is called the central paradox of sampling theory[14]], if one selects the members of the sample solely on the basis of one's own prejudices or ideas as to which particular members should be in the sample, and if exactly the same sample is chosen by the random process, the former sample would be inadmissible whereas the latter would be perfectly fine. Stuart says that this paradox is a hard pill to swallow. Nevertheless, he argues that the pill has to be swallowed in order to safeguard against unscrupulous researchers exercising their subjective biases. The notion of bias, incidentally, is different to that which we encounter in classical statistics with regards to the parameter estimation [cf. section 2.4.1]. Here the term 'bias' is used in synonymy with the term 'prejudice'.

[14] Stuart (1962)

So, from the point of view of a proponent of random sampling lack of randomisation opens judgement sampling to influence by biases or prejudice on behalf of researchers. However, there are several advantages that judgement sampling enjoys over random sampling. Judgement sampling focuses on the quality of the outcome of the procedure rather than on the procedure itself. This implies that the proponents of judgement sampling find it impossible to 'swallow' the paradox of random sampling. Judgement sampling is less costly and can be carried out much faster than random sampling. Judgement sampling avoids the issue of non-response. That is, situations when the individuals who have been painstakingly selected by random sampling cannot be reached or refuse to participate. Moreover, there is some inductive support for the effectiveness of representative sampling – for instance, success of political pre-election opinion polls, although it can be argued that the polls themselves lead to changes in behaviour on behalf of the electorate. Voters may engage in strategic voting on the basis of the results of such polls, thus creating a self-fulfilling prophecy.

On balance, it seems that a halfway house approach is desirable. That is, doing a stratified random sampling depending on the amount of knowledge in the field, costs/speed required. If a lot is known about the phenomenon, and costs of random sampling are prohibitive, then representative sampling is just the ticket.

1.2.2 Model Selection

Model selection of a certain kind forms a considerable part of this thesis. So, suppose we have gathered our sample in a way suitable for us. We now have several models/equations that could be candidates for *the* predictive model we are to use. On what basis are we to pick one?

The first choice that we have to make is whether we are to confirm/validate/test/choose between models that we have arrived at prior to considering our sample, or whether we are to attempt to construct the model by looking at the data – that is, by 'letting the data speak for themselves'. In this thesis we will be concerned with the former approach. There is a consensus that the latter methodology often leads to problems with spurious correlations and models that are

not useful for the purposes of prediction. More will be said of this topic in chapter 3, particularly in section 3.4.1.4.

The second choice is the choice of the method thereby the model is to be chosen. As it is already mentioned in section 1.1.2, chapter 2 is dedicated to elucidation of the traditional statistical approaches to this issue. However, the core of this thesis [i.e., chapters 3 to 5] is dedicated to considering more novel approaches.

1.2.3 Parameter Estimation

We saw in section 1.1.3 that in the Classical approach to statistics one uses estimators that satisfy the list of desirable properties, whereas in Bayesian statistics one does not focus on estimation as such, but when one does do estimation, then one usually uses estimators that provide maximum a posteriori probability of the model being correct. Consequently, given these different objectives, the estimators, and hence the estimates, often differ between these methods. As we stated previously, we consider parameter estimation only insofar as it is relevant to our central issue of model selection. The parameter estimation debate is tangential to the issue of model selection, so we will not be going into it in any detail. For some arguments within the parameter estimation debate, see Howson and Urbach (2006).

1.3 Probability Theory

1.3.1 Probability Primer

Statistical modelling is done in terms of random variables and probability distributions. These notions are part and parcel of the probability theory. So, in order to come to grips with how statistical modelling is done, we have to familiarise ourselves with central tenets of probability theory. This is the task of this section. There are several notions in this as well as in subsequent sections such as variance and statistical expectation, which are prima facie do not seem to do any useful work. However, familiarity with these formal tools is needed, for without it one would find it very difficult to comprehend the arguments given in later chapters of the thesis.

For the purposes of this section we introduce 'probability' as a primitive term [cf. Gillies (1973):232]. We shall not engage into the issue of interpretation of probabilities unless required for the discussion at hand[15].

The mathematical theory of probability can be thought of as a study of logical structure of uncertainty. This logical structure is determined by the axioms and all of their deductive consequences[16].

By way of introduction, let us consider a game of chance – for instance, that of the throwing of a die. What are the possible outcomes? We can get either 1, 2, 3, 4, 5 or 6. The set of these values constitutes the outcome space: {1, 2, 3, 4, 5, 6}, that is, the set of all possible mutually exclusive outcomes of the process. Each of these outcomes is called a basic event. An event is a set of basic events. An event occurs when one of the set of the constitutive basic events occurs. For example, in our case an event can be 'the number on the die is odd'. The corresponding set of the basic events is: {1, 3, 5}.

We can think of events as propositions that are closed under the truth-functional logical operators of conjunction, disjunction and negation. For our purposes the distinction between propositions and events is not important. We treat them as mutually substitutable.

Probability is measured by a real number between 0 and 1, where number 0 corresponds to a logical contradiction and 1 corresponds to a tautology[17].

[15] For comprehensive surveys of the issues involved in interpreting probabilities please see Gillies (2000) and Hájek (2009).

[16] The presentation of the probability theory in this section including the axioms thereof closely follows Howson and Urbach (2006): chapter 2.

[17] Also, it is important to note that formally, an impossible event has the probability of 0, and an event which is certain to occur has the probability of 1, but the converse does not hold in either case [cf. Kolmogorov (1956):5]. That is, if an event has zero probability of occurring, it does not imply that it is impossible to occur. Let us once more use the cannon ball example. What is the probability of the shot cannon ball landing *exactly* 125 metres away from the cannon? The answer is that it is zero, for we represent distance by real numbers, of which there are uncountably infinitely many. So the probability of picking one of them at random would be $1/\infty$ that is zero. But clearly the event of the cannon ball landing exactly 125 metres from the cannon is not impossible.

The intuitive idea of probability is formalised in terms of the following axioms [in what follows P(Y) stands for the probability of any event Y]:

(1) $0 \leq P(A) \leq 1$ for any event A in the domain of P
(2) P(logical truth) = 1
(3) P(A or B) = P(A) + P(B) for any mutually exclusive events A and B
(4) P(A|B) = P (A and B) / P(B) where P(B) > 0

Conditional probability P(A|B) is the probability of occurrence of event A given that event B has occurred. For example, suppose that event A is that a roll of the die results in number six, and B is that a roll of the die results in an even number. Supposing further that the die is fair [this is the case when, for instance, its centre of gravity lies in its geometrical centre], P(A) = 1/6. However, conditional on the die giving us an even number as the outcome, the probability of observing six is 1/3. That is, P(A|B) = 1/3.

Axiom 3 is sometimes extended to countably infinite sets of events mostly for mathematical convenience[18]. This, however, introduces some conceptual issues, but these need not concern us here[19].

Axiom 4 is sometimes introduced as a definition of conditional probability. However, we will treat it as a postulate on par with the other three. 'The reason for this is that in some interpretations of the calculus, independent meanings are given to conditional and unconditional probabilities, which means that (4) cannot be true simply by definition.' [Howson and Urbach (2006):16] Again, nothing in this thesis hangs on this point.

An important deductive consequence of the Axiom 4 is Bayes Theorem [its importance is discussed at length in section 4.1]. In its most commonly used form it is:

[18] 'For, in describing any observable random process we can obtain only finite fields of probability. Infinite fields of probability occur only as idealized models of real random processes. *We limit ourselves, arbitrarily, to only those models which satisfy [the Axiom of Countable Additivity]*. This limitation has been found expedient in researches of the most diverse sort.' [Kolmogorov (1956):15]
[19] For discussion cf. Gillies (2000):66-69, Howson (2008), Williamson (1999).

$P(A|B) = P(B|A)P(A) / P(B)$ where $P(B) \neq 0$

Also, often P(B) can be substituted by the expression which is called the total law of probability: P(B) = [P(B|A)*P(A) + P(B|notA)*P(notA)]

Let us illustrate the use of Bayes theorem with the following example. Suppose that we have a group of 100 students, 70 of whom study at college R and 30 are at college U. These students are to sit an examination, in which they either succeed or fail. Let us introduce the following propositions. J: A student studies at college R. C: A student studies at college U. S: A student passes the exam. F: A student fails the exam. Suppose further that we believe that a student from college R has 0.8 probability of passing the exam, whereas a student from college U has 0.4 probability of succeeding. We can represent these by conditional probabilities: P(S|J) = 0.8, P(S|C) = 0.4. Also P(J) = 0.3, P(C) = 0.7. Now suppose we would like to find out what the probability is of a student who passes the exam to have studied at college R [that is P(J|S)]. For this we employ the total probability form of Bayes theorem [since J and C are mutually exclusive exhaustive events]:

P(J|S) = P(S|J)*P(J)/[P(S|J)*P(J) + P(S|C)*P(C)] = 0.8*0.3/[0.8*0.3+0.4*0.7] = 0.24/0.52 = 0.46 [approximated to 2 decimal places – hereafter 2 d. p.]

An important concept in the theory of probability is that of the *probabilistic independence*. Events A and B are said to be probabilistically independent just in case P(A and B) = P(A)xP(B). Hence substituting this expression into the Axiom 4 [cf. page 23] we obtain the result that A and B are probabilistically independent if and only if P(A|B) = P(A) and P(B|A) = P(B).

An important concept in statistics is that of a random variable. A random variable is a mathematical function from the space of elementary events to the elements of the set of real numbers. For example, suppose that we roll a die twice and record the outcomes of both rolls. A random variable in this case could be a summation between the two outcomes. So, if the first throw yields 1, and the second throw 5, then the realised value of the random variable is 6. By convention we denote random

variables by capital letters, and particular numerical realisations thereof by small letters. We are interested in how probability values are distributed over every possible realisation of the random variable(-s). Various probability distribution models provide a summary of this information. A probability distribution model is a function that maps numerical values of random variables onto probability values. That is, a probability distribution model tells one what probability value is associated with each value of a given random variable. We will ordinarily refer to probability distribution models as just probability distributions, as it is conventionally done in statistics. The following table is an example of a probability distribution P(X) for the random variable in this paragraph [denoted here as X] on the assumption that every elementary event is as probable as every other elementary event:

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

A cumulative probability distribution shows the accumulation of probability up to a given value of the random variable. It is commonly denoted as F(X=x), where x is a particular realisation of the random variable as a result of the experiment. F(X=x) is the probability that X takes on a value smaller than or equal to x. E.g., in our case of the die throwing experiment, F(X=4) = P(X=2) + P(X=3) + P(X=4) = 6/36.

There is also a distinction between discrete and continuous probability distributions. In discrete distributions random variables can take on a finite or countably infinite number of values. So in our example with the rolling die we have a discrete probability distribution, since the random variable can only take discrete values. Let us illustrate the idea of a probability distribution with the example of the Binomial distribution, since it is reasonable to suppose that our experiment of throwing the die follows the Binomial distribution, which is a particular example of a discrete distribution.

The Binomial distribution applies in cases when there are just two exhaustive [that is, the sum of the probabilities of such events equals to one] and mutually exclusive [i.e., if one event takes place then the other event has the probability zero and vice versa] events. One event is usually referred to as 'success' and the other is 'failure'. If the experiment is repeated *n* times, with the repetitions being independent of one another, and if the probability of success *p* each time is the same, then
$P(x) = n!p^x(1-p)^{(n-x)}/x!(n-x)!$, where x is the variable that denotes the number of successful outcomes.

So, let us apply it to the die-throwing example. Suppose that we define event A as 'the number on the die is even' and event B – 'the number on the die is odd'. Events A and B are mutually exclusive – either A or B happen, but both of them cannot do, and exhaustive – every basic event belongs to either one[20]. Let us suppose that we are going to throw the die 10 times, and suppose that these throws will be independent of each other [that is, the probability of observing an even or odd number on each throw does not depend on the outcomes of the previous throws[21]]. Suppose further that each time we throw, the probability of A [let us call it 'success'] and B [call it 'failure'] is 0.5 respectively. Let us define a random variable X as representing the number of successes. So, for example, what is the probability that we will see exactly four even numbers?
$P(x = 4) = 10!\ 0.5^4 \times 0.5^6 / 4!6! = 0.205$ [correct to 3 decimal places]

Now, continuous probability distributions are such that the random variables that they cover can take on an uncountably infinite number of values. Some random variables are continuous, i.e., they belong to the set of real numbers rather than just integers, as it is the case for discrete distributions. E.g., X is the volume of milk that a herd of cows yields, or the temperature in a room at certain time. Also continuous distributions are often introduced for mathematical convenience. Continuous distributions have probability density function $f(x)$ such that $f(x) = dF(x)/dx$

[20] Thus, B is called a *complement* of A. I.e., A is logically equivalent to not B, and P(B) = 1 – P(A).
[21] A more rigorous definition of probabilistic independence is this: X is probabilistically independent of Y just in case P(X|Y) = P(X). That is, the probability associated with various values of random variable X stays the same whatever value random variable Y takes on.

The distributions of continuous variables are called probability density functions rather than just probability distributions because, for any point x, they show the probability of the random variable taking on a value in the region of x. We are referring to probabilities in the region of certain values rather than to probabilities of point values themselves, because every point value [of uncountably infinitely many point values] of a continuous random variable has probability zero. An example of a continuous distribution is the Normal distribution. We shall say more of Normal distribution later.

Another notion that we ought to introduce is that of statistical expectation. The expectation operator is quite convenient in order to define important properties of probability distributions. The expectation of a random variable is defined as a probability weighted average of all the values that the variable can take. That is, $E(X) = \Sigma^{n}_{i=1} x_i P(x_i)$, $i = 1, 2, 3, \ldots, n$ in the discrete case and $E(X) = \int xp(x)dx$ [on the range of values of X from the smallest to the largest] when X is a continuous variable. Intuitively E(X) can be thought of as an average value of the random variable in the long run.

Moments of a probability distribution are convenient ways to summarise some important properties of the distribution. We will concentrate on the two most important quantities – the mean μ and the variance σ^2. The mean is a measure of location. It shows us where the centre of the distribution is. The mean is equal to the expectation of the random variable. I.e., $\mu = E(X)$. For instance, consider the die rolling example with the discrete random variable X and the following probability distribution:

X_i	1	2	3	4	5	6
$P(X_i)$	1/6	1/6	1/6	1/6	1/6	1/6

In this case $\mu = E(X) = 1/6(1 + 2 + 3 + 4 + 5 + 6) = 21/6 = 3.5$ This of course does not mean that we would expect to obtain an outcome of 3.5 at some point in our experiment, for such an outcome is not in the set of possible values that our random variable can take on.

There are other measures of location such as the median [which refers to the middle value of the random variable rather than the probability-weighted average thereof], which have their advantages [the median is not affected by the outliers – the values that lie far away from the main body of data]. However, the mean is the predominant measure of location chiefly because it has 'nice' mathematical properties.

Another important moment of a probability distribution is variance [σ^2]. Variance is a measure of dispersion. It indicates how spread out the values of the random variable are around the mean of the distribution. In the discrete case, the variance is equal to the sum of the probability-weighted square deviations of every possible value of the random variable from the expectation of the random variable. In its simpler form, it can be demonstrated that the variance is equal to the expectation of the squared random variable minus the squared expectation of the variable itself. That is, Variance $(X) = E(X^2) – [E(X)]^2$. The standard deviation σ is the square root of the variance. The standard deviation is measured in the same units as the variable itself. Quite often the mean and variance are sufficient to uniquely define a probability distribution function [p.d.f.] – e.g., this is the case for normal distribution. In the die-rolling example above $\sigma^2 = 1/6(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) – 3.5^2 = 81/6 – 12.25 = 13.5 – 12.25 = 1.25$. Thus $\sigma = 1.118$[22] (3 d. p.)

1.3.2 Normal Distribution

Quite possibly the most important probability distribution that is used in statistics is the so-called Normal Distribution. Firstly, a lot of phenomena have been observed [at least approximately] to follow this distribution – e.g., distribution of heights, exam marks, etc. Secondly, the importance of Normal Distribution stems from the Central Limit Theorem – if the individual observations constituting a sample are independently identically distributed, and as the number of such observations becomes large, the sample mean tends to be normally distributed, irrespective of the form of the distribution of the population itself, as long as the population variance is finite. An implication of the Central Limit Theorem [CLT] is that the binomial

[22] Neither variance nor standard deviation can take on negative values.

distribution becomes approximately Normal as the number of observations n increases (some authors claim that in practice the approximation is reasonably close once n > 50).

A continuous random variable X is said to be normally distributed with mean μ (E(X) = μ) and variance σ^2 (Var(X) = σ^2) [ordinarily represented as X ~ N(μ, σ^2)] when its probability density function (PDF) is $f(x) = e^{-(x-\mu)\,(x-\mu)/2\sigma\sigma} / (2\pi\sigma^2)^{1/2}$

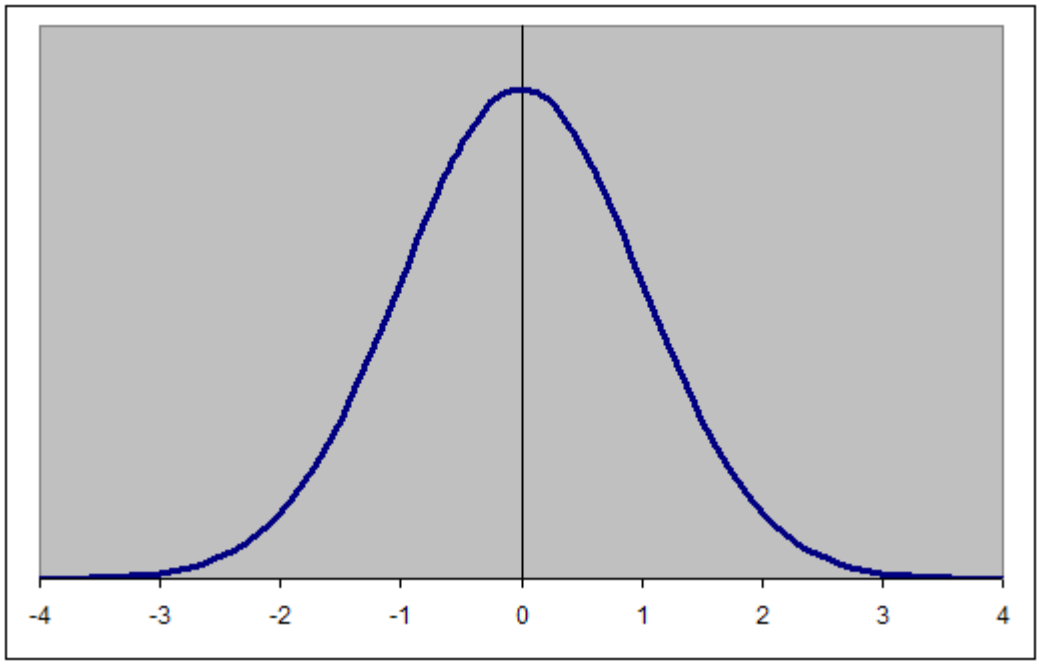

Diagram 1

The picture above represents the p.d.f. of the standard normal distribution. That is, the distribution of a random variable Z such that Z ~ N(0,1). We measure the p.d.f. of Z along the vertical axis, and the values of Z along the horizontal one. The standard normal distribution has practical importance because it has been tabulated. Any linear combination of the normally distributed random variable is itself normal, so if X ~ N(μ, σ^2), then Z = [(X – μ) / σ] ~ N (0,1). Hence, any normal distribution can be transformed into the standard normal for ease of calculations. For example, suppose that our random variable X represents the heights of males within the London Borough of Waltham Forest [recall the example used in the beginning of this chapter]. Suppose further that X is normally distributed with E(X) = 180 cm, and σ = 10 cm. That is, X ~ N(180, 10). Given this set up, suppose that we would like to find

out what the probability of observing at random a male who is taller than 195 cm is. In order to do so, we convert our random variable X into Z thus:
$P(X > 195) = P(Z > (195 – 180)/10) = P(Z > 1.5)$ In order to calculate this value by ourselves we would need to integrate the probability density function between 1.5 and infinity. This is quite a laborious task, so instead we can look the value up in the tables for standard normal[23]. From such tables, $P(Z > 1.5) = 0.0668$.

Now, let us pause for a short while to see what we have done so far and where are going to in the rest of the chapter, and, indeed, in the rest of the thesis. To begin with, we considered the issue of scientific prediction, restricting our attention to statistical problems. The example that we used was that of wishing to predict the weight of a male who resides within the London Borough of Waltham Forest on the basis of his height. We saw that in order to get to grips with a problem of this sort we needed to gather a suitable sample of observations, select a statistical model and estimate the parameters within the chosen model. We said that we would restrict our attention to the issue of model selection. Since statistical models are formalised in terms of probabilities, probability distributions and their moments, etc., we overviewed the probability theory and the necessary concepts and terms which equipped us to understand how statistical models work. In chapter 2 and in section 4.1 we consider the two methodologies that currently dominate the field of statistical reasoning, viz., Classical and Bayesian respectively. We consider the issues that each of these methodologies has. From the third chapter onwards the thesis is dedicated to two major alternative approaches to model selection which have been developed since the early 70s of the 20th century, viz., Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC). The chapter five considers the putative philosophical consequences of the model selection methodologies.

[23] These tables are widely available. E.g. http://www.math.unb.ca/~knight/utility/NormTble.htm

2. Classical Statistics

The name Classical statistics is, strictly speaking, a misnomer. Rather than being a unified methodology, it is in fact a heterogeneous collection of various methods such as R. A. Fisher's, Neyman-Pearson's, parameter and confidence interval estimation techniques, etc. However, we will follow the numerous text books on practical application of statistics in using this somewhat misguided terminology as a convenient umbrella term in cases when it does not matter which particular technique or method within it we refer to.[24]

At the outset of the expositions of Classical statistics in this section, let us note a salient distinction between uses of probability between the Classical and Bayesian schools of statistical thought. In the latter, '…probability is used to provide a post-data assignment of degree of probability, confirmation, support or belief in a hypothesis…', whereas in the former '…probability is used to access the probativeness, reliability, trustworthiness, or severity of a test or inference procedure.' [Mayo (2005):803] Simply put, in Bayesian statistics probability applies to hypotheses and data whereas in Classical statistics probabilities are used for assessment of inference procedures themselves. In other words, in Bayesian statistics hypotheses have probabilities whereas in the Classical context probabilities are used to control of various types of errors given inference procedures may generate. Note, incidentally, that we use the terms 'hypothesis' and 'model' interchangeably.

2.1 Fisher[25]

The modern approach to statistical inference was started by R. A. Fisher [Mayo (2005):804]. He considered that

> '…the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.

[24] Classical statistics is also often referred to as Frequentist due to the eponymous interpretation of probability that these methods usually use.
[25] The exposition of Fisherian and Neyman-Pearson methodologies closely follows Royall (1997) and Newbold (1995).

> The problems which arise in reduction of data may be conveniently divided into three types:-
> (1) Problems of Specification. These arise in the choice of the mathematical form of the population.
> (2) Problems of Estimation. These involve the choice of methods of calculating from a sample statistical derivatives, or as we shall call them statistics, which are designed to estimate the values of the parameters of the hypothetical population.
> (3) Problems of Distribution. These include discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known.
> As regards problems of specification, these are entirely a matter for the practical statistician, for those cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type. In other cases we may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested *a posteriori*. We must confine ourselves to those forms which we know how to handle, or for which any tables which may be necessary have been constructed. More or less elaborate forms will be suitable according to the volume of the data. Evidently these are considerations the nature of which may change greatly during the work of a single generation.'
> Fisher (1922):311, 313, 314

It does show that Fisher thought that the problems of model selection [or as he referred to them as problems of model specification] are important. However, in his methodology he confined himself to problems of estimation and distribution. On the other hand, Fisher's method can still be considered to constitute model selection in the sense that in it we test an element of a given model, and if it is deemed to be incompatible with data, we then are faced with the choice to either choose a different element of the same model, or indeed to choose a different model – that would presumably be 'a matter for the practical statistician'. But we are getting somewhat ahead of ourselves.

Let us explain Fisherian methodology by means of an example. Suppose that we have a die-rolling set up such that the random variable A represents the number of even outcomes of rolling the die. We would like to provide a statistical model for this set up. In Fisher's methodology one hypothesises a single model [referred to as the 'null hypothesis'] with fixed values of parameters. In our case this idea corresponds to us hypothesising that, for instance, the phenomenon follows binomial distribution with success parameter $p = 0.5$ [let us define a successful outcome as such that when we observe an even number of dots on the die] corresponding to our

supposition that the die is fair. So, our model is that A is binomially distributed with p = 0.5. In order to complete the model, we also have to decide how many observations our sample is to consist of. Suppose for the sake of argument that we set out to roll the die 120 times. Hence A is binomially distributed with n = 120 and p = 0.5. Then we observe a relevant sample of data. In our case, such a sample would have 120 throws of the die, with the outcomes being either even or odd numbers on the upper most surface of the die when it comes to rest. It is standard practice to approximate binomial distribution by means of a normal distribution. The main conditions for doing so are that n is sufficiently large [most authors in statistical literature consider $n > 50$ as large enough] and that p is not too close to either 0 or 1. Both of these conditions obtain in our case, so the use of normal approximation is warranted. Let us put in some numbers for ease of understanding. Suppose that we roll the die 120 times[26], and that 70 times it gave us an even number and 50 times an odd one. We will use the Normal approximation to the Binomial, where the mean and variance are calculated thus: $\mu = np$, $\sigma^2 = np(1-p)$. Hence, A is normally distributed with the mean $\mu = 120x0.5 = 60$ and variance $\sigma^2 = 120x0.5x0.5 = 30$, in short: $A \sim N(60, 30)$

Now, let us distinguish two sub-methods within Fisher's methodology according to which we can proceed from here to test our supposition that the die is fair. The first one is the method of rejection trials, the second is the method of calculating so-called *P*-values.

2.1.1 Rejection Trials

So, we have our null hypothesis – that is, the model with the values of parameters fixed. In rejection trials the idea is that we test our model against data. The idea of testing is that one checks one's sample of data against one's model to see whether the data are consistent or significantly inconsistent[27] with the correctness of the model [indeed, this facet of Fisherian methodology is often referred to as a 'test of

[26] We assume that each throw is independent and identically distributed, that is, each throws follows the Binomial distribution where the probability of success is constant and the same for each throw.

[27] The concept of *significant* inconsistency may strike the reader as odd, for in, for example, propositional logic the concept of consistency is binary – either a set of propositions is consistent or it is not. We will look into this Fisherian use of the concept later in this section.

significance']. Our model specifies a probability of observing every possible sample. We set a threshold probability value [it is usually referred to as the *level of significance*] and devise the following decision rule. If our model specifies that the observed outcome or outcomes at least as extreme have the probability of occurring greater than the critical value, we do not reject our model and tentatively uphold it until the next test. By 'outcomes that are at least as extreme' we mean those outcomes, that under the assumption that the null hypothesis is true have the probability that is at most as large as that of the actually observed outcome. If, on the other hand, the observed outcome or outcomes that are 'at least as extreme' have, according to our model, the probability of occurring which is lower than the level of significance, then we reject the model and seek an alternative one. With the level of significance and decision rule in place, we observe the sample, and comply with our decision rule. In order to make this method clearer, let us carry on with our example.

So, in our die rolling example we have binomial set up with n = 120 and p = 0.5, which we approximate by $A \sim N(60, 30)$. Suppose [as it is commonly done] that we set the level of significance at 0.05. Testing at this level of significance has become conventional, although some practitioners prefer 0.01 or other levels – the choice of the level of significance appears arbitrary[28]. We now carry out our experiment and suppose that we observe 70 even numbers and 50 odd numbers respectively. Now, the question is: how likely are we to observe this outcome or the outcomes that are at least as extreme under our hypothesis of the fairness of the die? In our example the outcomes at least as extreme are: 71 even and 49 odd, 72 even and 48 odd, and so on, as well as 50 even and 70 odd, 49 even, and so on, because observing 50 even and 70 odd has the same probability as that of observing 70 even and 50 odd due to the symmetry of the distribution around its mean value, which in our case is 60 even and 60 odd; and 49 even, 48 even and so on all have lower probability of occurring than 50 even and consequently than 70 even. So, due to the symmetry, $P(A \geq 70) = P(A \leq 50)$. Using the transformation of our normal distribution into the standard

[28] '[Fisher] advocated 5% as the standard level (with 1% as a more stringent alternative); through applying this new methodology to a variety of practical examples, he established it is a highly popular statistical approach for many fields of science. … [Fisher] also wrote that "it is usual and convenient for experimenters to take 5 percent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard…" [Lehmann (1993):1243, 1244] For a contemporary debate on this topic see Hoover and Siegler (2008) and references therein.

normal, and the tabulation of the standard normal distribution, that we familiarised ourselves with previously, we obtain the following:

$P(A \geq 70)$ or $P(A \leq 50) = 2p(Z \geq (70-60)/30^{1/2}) = 2p(Z \geq 1.83) = 2(1 - p(Z \leq 1.83)) = 2(1 - 0.9664) = 2x0.0336 = 0.0672$

That is, the probability of obtaining 70 even numbers out of 120 throws or an outcome that is at least as improbable is 0.0672 [that is, 6.72%] on the hypothesis that the die is fair. Since this probability is greater than our pre-determined rejection threshold value of 0.05, we do not reject our hypothesis of fairness of the die at 5 per cent significance level. Note, however, that if our significance level was, say, 0.1, we would have rejected the null hypothesis[29].

To clarify, the reasoning here is roughly this: we should reject a hypothesis upon observing an outcome [in our example that is 70 out of 120 throws] such that the probability of observing this or outcomes at least as extreme on supposition that the hypothesis is true is 'low' relative to the probability of observing other possible outcomes of the experiment. The probability is deemed 'low' when it is below the significance level [here it is 5%]. So our particular model has survived this test.

2.1.2 *P*-values

The method of p-values is formally very similar to that of rejection trials. The difference lies predominantly in the interpretation of results.

The p-value is the probability of obtaining an outcome or a more extreme one on the supposition that the hypothesis is true. Recalling the example that we used in the

[29] 'Another consideration that may enter into the specification of a significance level is the attitude toward the hypothesis before the experiment is performed. If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low. (A low significance level results in the hypothesis being rejected only for a set of values of the observations whose total probability under the hypothesis is small, so that such values would be most unlikely to occur if [the null hypothesis] were true.)' Lehmann (1986):70 It seems that in such cases Lehmann advocates using the significance level of something like 0.01. However, motivating such choice by 'firm belief that the hypothesis is true' does not seem to be open to classical statisticians, for they would need to explain further what constitutes this 'firm belief' [since they deny that hypotheses have *probabilities* of being true – see page 31], whereas this has a natural interpretation within the Bayesian statistics as there being a high prior probability of truth of the null hypothesis – cf. section 4.1.

rejection trials subsection, the p-value there was 6.72%. However, rather than creating a rule which directs us to a decision as to whether to reject or not to reject the hypothesis at the pre-set level of significance, the p-value is taken to signify the strength of evidence against the hypothesis. This is based on the so-called Law of Improbability [here is a somewhat naïve rendition of it]: If the hypothesis implies that the probability *p* of observing a certain outcome is small, and the outcome has been observed, then *p* is evidence against the hypothesis, and the lower the numerical value of *p* the stronger this evidence is.[30]

There are several difficulties that Fisherian method runs into. Let us consider some of them.

Firstly, as we already mentioned, there is arbitrariness in choice of the significance level, so that one and the same observation may lead to either rejection or not of one and the same null hypothesis depending on that level. To be fair, this criticism only applies to the rejection trial method and not to the method of p-values.

Secondly, the accept/reject nature of the rejection trials method does not take into account the strength of evidence that the sample provides us with. Again, this is prima facie problematic for the rejection trials method, not for the p-values.

Thirdly, another issue with the rejection trials method is in the doubtful nature of the concept of what we call significant inconsistency, since in formal logic the concept of consistency is binary – for example, a set of propositions is either consistent or inconsistent. Fisher argues: '[Tests of significance] could 'disprove' a theory … and … when used accurately, [they] are capable of rejecting or invalidating hypotheses, in so far as these are *contradicted by the data*.' [quoted in and added italics by Howson and Urbach (2006):150] It is rather clear that the data with a low probability of occurring under the null cannot be *logically* inconsistent with it. The quote indicates that Fisher wants significant inconsistency to be as close as possible to logical inconsistency. Elsewhere Fisher (1956):39 equates his notion of statistical significance with the following disjunction: either the hypothesis is false or a very

[30] For an in-depth analysis cf. Royall (1997):chapter 3.

rare event has occurred. Practitioners typically supplement this notion of statistical significance with that of *practical* significance. For instance, Agresti and Finlay (2009):163 discuss an example of testing the hypothesis that on average the population of the USA holds moderate ideological views. That is, the hypothesis that the sampling distribution is normal [with the variance estimated from the sample] and that the population mean is 4 as measured on the ordinal scale from 1 representing extremely liberal views to 7 representing extremely conservative views. Supposing that in a very large sample the sample mean is 4.08, Agresti and Finlay (ibid.) calculate the p-value of approximately 10^{-11}, which is extremely statistically significant. However, they contend that in this context the difference between 4 and 4.08 is of no practical significance.

Fourthly, there are no alternative hypotheses provided, so that even if we do not reject the null, perhaps there is at least one other hypothesis out there that we also would not reject, and which perhaps would have a higher p-value indicating that there are more evidence against the null, so that the alternative is somewhat better. I.e., a hypothesis whose parameters were fixed at the values which turned out to be closer to the actual observations; in our example one such hypothesis would be a model with the success rate set at p=0.55 rather than p=0.5, as it was the case for the null.

Finally, even though the strength of evidence is attempted to be captured with the notion of p-values, the numerical expressions of p-values depend on how we define the outcome space, and as such they are arbitrary. Recall that in the example in this subsection we hypothesised that A ~ N (60, 30), and that we observed A = 70. Since the p-value is the probability of observing the actual outcome or outcomes at least as extreme on the supposition that the hypothesis is correct, we calculated the p-value as 0.0672. Now, for the sake of the argument, suppose that we have a colleague who is interested in our experiment[31]. Suppose that the colleague resides very far away from us, and that we have only the most primitive means of communicating with her. Knowing that we can only send her a signal in the form of a 'Yes' or 'No', we happened to have agreed with the colleague [when we had got a rare opportunity to

[31] This example is a modified version of the one used in Royall (1997):68.

meet her a long while ago] that we would communicate 'Yes' if we got A = 70 and 'No' if we got any other value. Hence, her outcome space consists of two values, viz., {70, not-70}, whereas ours is made up of 121 values. Now, our colleague also uses the Fisherian method, and wishes to calculate the p-value. Since A = 70 is *the* most extreme outcome that she can observe, her p-value is P(A=70) = 0.0138 [4 d. p.][32]. Our p-values differ whereas we observed the same evidence – 70 even numbers out of 120 throws of the die. This example illustrates the point that p-values depend on outcomes that did not happen. As Jeffreys eloquently puts it in a much-quoted passage [where '*P* [integral]' stands for 'p-value' and 'law' stands for 'hypothesis']:

> If *P* is small, that means that there have been unexpectedly large departures from prediction. But why should these be stated in terms of *P*? The latter gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution from the actual value is nearly always negligible. *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it. The same applies to all the current significance tests based on *P* integrals.'
> Jeffreys (1961):385

Arguably, the strength of evidence for or against any hypothesis should be solely based on the observations that have actually been made, and not on something that has never been observed. On this view, the way that one defines the outcome space should be irrelevant. The example that we use could hardly happen in modern academic life. However, this does not negate the methodological point it raises.

We believe that the discussion in this subsection have served to indicate that there are substantial issues with using Fisherian methods for choosing either a family of models or a particular model.

2.2 Neyman-Pearson

In the previous section we looked at Fisherian methodology. At the end of that subsection we noted several disadvantages that the methodology has. In order to

[32] We performed the calculation for P(A=70) using the binomial formula directly rather than the normal approximation, because when A is a continuous variable, any particular point value of it has the probability of zero.

overcome some of these disadvantages, J. Neyman and E. S. Pearson devise methodology that we are going to consider in this subsection.

In Neyman-Pearson hypothesis testing approach one postulates two hypotheses [rather than one as in Fisher's case], which are normally called the null hypothesis H_0 and the alternative hypothesis H_1. These hypotheses normally take one of the following three forms. First, both H_0 and H_1 are point hypotheses [that is, they are single models with different fixed values of parameters]. Second, H_0 is a point hypothesis and H_1 is a composite hypothesis [that is, a proper subset of a model with more than one element in it]. Third, both H_0 and H_1 are composite hypotheses. Having set up the hypotheses, one works out what the so-called rejection region is. The rejection region is calculated according to what is called the *Fundamental Lemma* by satisfying the following inequality: P(observation under H_0)/P(observation under H_1) $\leq k$, where k is a constant depending on both the significance level [the same concept as in the Fisherian methodology above] and the hypotheses themselves [see Howson and Urbach (2006):148]. Informally, this guarantees that the rejection region lies between H_0 and H_1. After that one observes the data. Then one follows this decision rule, on the crucial assumption that one of the hypotheses is true: if the data is in the rejection region, then the null hypothesis is rejected and the alternative hypothesis is accepted; if the data does not fall within the rejection region, then the null hypothesis is accepted. This approach is alternatively called error probabilistic, because one of the most crucial elements of this method is the control of error probabilities. There are two types of errors that can be committed. The null hypothesis is rejected whereas it is true [this is called a Type I error], and the null hypothesis is accepted whereas the alternative hypothesis is true [Type II error].

A salient analogy here is that of court trials. The null hypothesis there is the innocence of the defendant [presumption of innocence]. If the court convicts the defendant when she is innocent that is a type I error, whereas when the court of law pronounces the defendant innocent while she is guilty that is the type II error. It is important which hypothesis is considered to be the null and alternative [just like in the court example]. This is because sometimes the inference changes if the null and alternative hypotheses are changed around. We will show an example of this later in

this section. Ordinarily the reason that is given for non-arbitrariness of such a choice is that it is usually quite clear what is the natural choice as to which hypothesis should be the null and which should be the alternative one. The null is usually the default 'sceptical' hypothesis. E.g., at the drug trial one would naturally want the hypothesis that the given tested drug has no effect to be the null and the hypothesis that the drug has a positive effect to be the alternative. We only want to accept drugs when we are quite confident that they do have an effect. In this case the type I error would be to accept the drug as effective whereas it actually has no effect. The type II error would be to accept the notion that the drug has no effect whereas it actually has a positive effect. Sometimes it is not that clear what should be the null and what should be the alternative. We will consider as to why this may matter by using an example further in this subsection.

In the Neyman-Pearson [NP] approach one calculates the probabilities of committing each type of error. The prescription then is to try to minimise both error probabilities as much as possible. It is impossible to achieve these two objectives simultaneously[33]. For a given number of observations reduction in type I error implies increase in the type II error. So what normally happens is that the type I error is fixed at a desirable level [this level is usually called the critical level α, and is usually set anywhere between 10% and 1%] and then the required power of the test [power = 1 – P(type II error)] is achieved by increasing the sample size. The power of the test is the probability that a false null hypothesis is rejected.

For the purposes of illustrating the idea of Neyman-Pearson testing, to begin with we take the most simple example of testing two point hypotheses. That is, both the null and the alternative hypothesise that the phenomenon in question follows the respective probability distribution models, and that the relevant parameters have sharp values.

As it has become customary by now, suppose that we have a die, and that we have two alternative ideas as for the probability of obtaining even numbers when we throw the die [call it the rate of success]. Just as in the subsection on Fisher, we

[33] This is the case when the number of observations is fixed. However, both types of error can be reduced if the sample size is increased.

suppose that we are in the binomial set up with the throws of the die assumed to be independently and identically distributed – so binomial probability distribution models represent both null and alternative hypothesis. Now, suppose for the sake of clarity of exposition that our null hypothesis is that the success rate is 0.55, and the alternative is 2/3. We set out to test these hypotheses by throwing the die 120 times. Since the number of observations is quite large, we will be using the normal approximation to the binomial for mathematical convenience.

So, under the null hypothesis on average we expect to observe 66 even numbers, and under the alternative hypothesis we expect to observe 80 even numbers. We set the probability of type I error at 5%, which is the standard practice in classical statistics. Suppose that our experiment yields 70 even numbers [denoted as $\bar{X} = 70$]. Then let us calculate the minimum number of even numbers that we need to observe in order to reject the null. As before, under our null hypothesis the variance is $npx(1-p)$ = 120x0.55x0.45 = 29.7. Using this variance in the standard normal calculation, we obtain the following:

P (Z > 1.64) = 0.05 => P (($(\bar{X} - 66) / 29.7^{1/2}$) > 1.64) = 0.05 => P ($\bar{X}$ > 66+1.64x $29.7^{1/2}$)) = 0.05 => $\bar{X}$ > 74.94 [2 d. p.]

So, in order to reject the null hypothesis in favour of the alternative at α = 5% probability of type I error [it is also called the level of significance], we need to observe at least 75 even outcomes out of 120 rolls of the die. Since we actually observed 70, we do accept the null in this case.

Now, let us work out the probability of type II error, i.e., of accepting a false null hypothesis. In our binary set up, falsehood of null implies the truth of the alternative hypothesis. We have just established that we do not reject the null if we observe the number of evens to be less than 75. Probability of type II error then is the probability of observing less than 75 evens given that the true success rate is 80. Calculating the variance under the alternative hypothesis as $np(1-p)$ = 120x2/3x1/3 ≈ 26.67 (2 d. p.) and transforming into units of the standard normal and using the tables for it, we obtain:

$(75 - 80) / 26.67^{1/2}$ = -0.9682 So, P (Z < -0.9682) ≈ 0.1664

Consequently, the probability of rejecting a false null hypothesis in this test [that is, the power of the test] is 1 - 0.1814 = 0.8336.

Now, let us look at testing a point null hypothesis versus a composite alternative, and also composite null versus composite alternative. The former in our example above would be something like H_0: $p = 0.55$ vs H_1: $p > 0.55$. The latter: H_0: $p \leq 0.55$ vs H_1: $p > 0.55$. Notice that the assumption of the truth of either H_0 or H_1 becomes progressively more legitimate, particularly in the composite H_0 vs composite H_1 case, where this assumption is correct providing that we have selection the correct model. It is interesting to note that in both of these cases the answer is the same as it was in the point H_0 vs point H_1 case above – we would reject H_0 at $\alpha = 5\%$ just in case we observe 75 or more even out of 120. However, now we cannot calculate the power of these tests, because in order to do so H_1 has to specify particular point values for the parameters. So, what then of the idea that we should maximise the power at the given level of significance α? In these cases Neyman and Pearson employ the concept of Uniformly Most Powerful Unbiased (UMPU) tests. A test is Uniformly Most Powerful when for every model within H_1 the power is maximised. It is also Unbiased when for each model within H_1 the power of the test is not smaller than the significance level. Otherwise such a test would have a higher probability of rejecting a true H_0 rather than rejecting a false one, which Neyman and Pearson deem undesirable. Both of our tests above are UMPU. The idea of the UMPU test becomes clearer when one considers tests of this type: H_0: $p = 0.55$ vs H_1: $p \neq 0.55$. We will look into this important case in section 1.4.5.

Here is a summary of some salient features of the power of a test from Newbold (1995):371:

1. 'Everything else being equal, the farther the true mean μ_1 from the hypothesized mean μ_0, the greater the power of the test.
2. Everything else being equal, the smaller the significance level of the test, the smaller the power. In other words, reducing the probability of a Type I error will increase the probability of a Type II error.
3. Everything else being equal, the larger the population variance, the lower the power of the test. We are less likely to detect small departures from the hypothesized mean when there is greater variability in the population.

4. Everything else being equal, the larger the sample size, the greater the power of the test. Again, this is intuitively plausible. The more information obtained from the population, the greater the chance of detecting any departure from the null hypothesis.'

Having considered the Neyman-Pearson approach in some detail, let us identify some key shortcomings that the method has.

Firstly, the binary accept/reject set up is rather crude. The prescription to behave as if the accepted hypothesis was true [until further tests are carried out, that is] does not provide us with the information as to what amount of evidential support the hypothesis enjoys, or what amount of confidence we have in the truth of the hypothesis. It has to be stressed, however, that there is much disagreement on this point in philosophy of statistics. The proponents of the Neyman-Pearson methodology consider the binary nature of this approach as its strength. It allows them to answer the question 'What should we do, given the data?' rather than 'How should we interpret the data as evidence regarding a hypothesis or one hypothesis versus another?' [Royall (1997):4] Still, arguably we would be much more cautious with regards to decisions that we make on the basis of a weakly supported hypothesis [or, alternatively if we do not have a great amount of confidence in the truth of the hypothesis] rather than if the hypothesis had more evidence indicating its truth. On the other hand, error probabilities carry out this function indirectly. However, probability of type II error [that is, of accepting a false null hypothesis] crucially depends on what one chooses as the alternative hypothesis. As stated in point 1 above, the further the alternative hypothesis away from the null, the smaller the probability of type II error is [and, consequently the greater is the power of the test]. For instance, if in the example that we used in this section our alternative hypothesis was p = 0.7 rather than 2/3, then probability of type II error would have been approximately 0.0365 rather than 0.1664[34].

Secondly, there is arbitrariness in the choice of the level of significance and in the choice as to which hypothesis is the null and which is the alternative one. The issue

[34] The critical value: $(75 - 84)/(120 \times 0.7 \times 0.3)^{0.5} = -1.7928$. So $P(Z<-1.7928) \approx 0.0365$

with the choice of the level of significance is very similar to the issue with the rejection trials method within Fisherian methodology. The problem with the choice of the null and alternative is the following. Suppose that in the example that we used in this subsection we choose the null hypothesis p = 2/3 and the alternative p = 0.55. Then for the rejection region with α = 0.05: P (Z<-1.64) = 0.05 => P (($\bar{X}$ – 80)/$26.67^{0.5}$ < -1.64) = 0.05 => $\bar{X}$ < 80 - 8.4695 => $\bar{X}$ < 71.5305. So if we observe the value of $\bar{X}$ of 71 or smaller, then we reject p = 2/3 hypothesis in favour of p = 0.55. Suppose that we observe $\bar{X}$ = 73. Under this set up we would accept the p = 2/3 hypothesis whereas originally given this observation we would accept the p = 0.55 hypothesis! Notice also that, unlike the cases of a court trial or test of a new drug [where it is claimed that the default position of the presumption of innocence or the hypothesis of the drug having no effect respectively both naturally play the roles of the null hypothesis], there is no obvious reason in this case as to why one of these hypothesis should be the null.

Thirdly, the approach suffers from something called Lindley Paradox. In fact Fisherian approach has the same issue. According to Lindley Paradox, as the number of observations grows, the proportion of successes at which we would just reject the H_0 at a given level of significance becomes arbitrarily close to the proportion stipulated by null, and the power of the test tends to one. So even a tiny deviation of the proportion in sample from that of H_0 is sufficient to reject the H_0, which is counter-intuitive. Here is an illustration using our example of testing H_0: p = 0.55 vs H_1: p = 2/3. Previously we noted that we would reject H_0 at α = 5% if we observed 75 or more evens out of 120. That is, if the proportion of evens in the sample was greater than 74.94/120 = 0.6245. The power of the test was 0.8336. Now, suppose that our sample consists of 12000 observations. Then for the rejection region: P (Z > 1.64) = 0.05 => P ((($\bar{X}$ – 0.55x12000) / $(0.55x0.45x12000)^{1/2}$) > 1.64) = 0.05 => P ($\bar{X}$ >6600+1.64x $29700^{1/2}$)) = 0.05 => $\bar{X}$ > 6689.38 [2 d. p.], which corresponds to observing the proportion of 0.5574 or more of evens in the sample rather than 0.6245 when the sample had 120 observations. In order to calculate the power of this test we require the following quantity: (6689 – 8000) / $2667^{1/2}$ = -25.36. So, P (Type II error) = P (Z < -25.36) ≈ 0 Hence the power of the test is approximately 1. To counterbalance this counter-intuitive result classical statisticians generally advise to

reduce the level of significance as the number of observations grows so that the test becomes less sensitive to small differences. However, this is a rather *ad hoc* manoeuvre that has no clear rationale with the NP methodology. Nonetheless, practical experience in using the NP framework suggests the expedience of this move in order to align NP with the intuitions of its practitioners.

2.3 Fisher vs Neyman-Pearson

In Neyman-Pearson approach to hypothesis testing to 'accept' the null hypothesis means that '…the data available do not provide enough evidence for rejection of the null hypothesis, given that we want to fix at alpha the probability of rejecting a null hypothesis that is true.'[35]

So, what does this methodology prescribe that we do with regards to selecting a probability distribution model? If we accept the null hypothesis in the sense given above, then we should behave as if the null hypothesis were true. However, if we reject the null hypothesis then Neyman and Pearson urge us to behave as if the alternative hypothesis is true. They call this approach 'inductive behaviour'.

In contrast, Fisher's rejection trials are very much like Karl Popper's Falsificationism. Here is an unsophisticated rendition of Falsificationism. Scientists entertain certain hypotheses [conjectures]. There is no amount of evidence that would establish a given hypothesis as true [cf. the well-known problem of induction]. However, a single observation that is logically inconsistent with the hypothesis shows it to be false. So, rather than confirming hypotheses what one ought to do is to try to disconfirm [i.e., falsify] them. Similarly, in the rejection trials method, one sets up a structure akin to modus tollens. To repeat the discussion in section 2.1, if the hypothesis is true, then the given observation [in statistics usually a set of observations – a sample] has a certain probability of being observed. However, if the probability of observed sample is below a pre-determined threshold [i.e., the level of significance], then the observations are deemed to be significantly inconsistent with the hypothesis [in the sense that they are too improbable under the

[35] Newbold (1995):329

hypothesis], hence the hypothesis is rejected. If, however, the probability is not below the threshold value, then the hypothesis is left in use until next trial. Fisher referred to this as 'inductive inference', which seems unwarranted because Popper considered that by his methodology of falsification he 'dissolved' the problem of induction. So if Fisher's method is statistical falsificationism, then, presumably there is also no induction taking place, but at best corroboration through survival of numerous tests. However, the p-value methodology, which takes the p-value to measure the strength of evidence against the hypothesis [the lower the p-value the stronger is the evidence against the hypothesis], does not seem to align with falsificationism. We shall consider other ideas of Popper in chapters 4 and 5.

Neyman and Pearson [NP] considered their methodology to be an improvement on Fisher's, in that they introduced the idea of the power of the test. From their perspective the p-values only measure the probability of rejecting a true null hypothesis. However, in Fisher's method there were no way to control for the probability of accepting the false null, which the power allows one to do. Fisher thought that the power should be a qualitative notion, for its quantitative calculation often involves unknown alternative hypothesis [such as is the case when the alternative is composite]. There are more points of disagreement between Fisher and NP, but there are of no consequence for our purposes. Interested reader is referred to Lehmann (1993) and Lenhard (2006).

Indeed, notwithstanding the issues from which the NP method suffers, it can be considered to be a proper method of model selection. There are two competing hypotheses. Of course traditionally the hypotheses have the same mathematical structure, but this is not a necessary attribute of the method. Also, importantly, in the case when we test point [simple] H_0 versus composite H_1, the null hypothesis has not adjustable parameters and the alternative hypothesis has one adjustable parameter. This notion is going to be discussed at length in chapter 3. It will for now be sufficient to say that the difference in the number of freely adjustable parameters is an essential part of the model selection methods discussed in chapters 3 and 4. Finally, the Neyman-Pearson point null versus composite alternative case shall be used in chapter 5.

2.4 Point Estimation

In sections 2.1 – 2.3 we looked at the classical methods of hypothesis testing. However, instead of testing hypotheses, scientists sometimes require estimates of parameter values from data. So, sections 2.4 and 2.5 are dedicated to brief introductions to the classical techniques of estimation by point values and intervals respectively, familiarity with which shall be useful for understanding the material in the subsequent chapters.

2.4.1 Properties of Estimators

Suppose that, rather than test hypotheses with regards to the probability distribution and the value of the population parameter as was done in the previous subsection, we would like to *estimate* a population parameter on the basis of a sample that we have drawn from the population and on the assumption that we have the right model-type. For instance, we may know that our population of interest is normal and may know the value of the standard deviation, but not know the value of the mean. In this case we come up with an estimator. That is, a function that has the values of sample observation as its inputs and the estimate of the relevant parameter as the output. How do we come up with such a function? After all, we can think of many possible estimators. In classical statistics the estimators have to have desirable properties, i.e., unbiasedness, consistency and efficiency[36]. In our case, it seems natural to estimate the population mean by the sample mean $\bar{X}$. The sample mean does possess the desirable properties.[37]

2.4.2 Mean Squared Error

Suppose that we have two estimators such that the first one is unbiased but it has a relatively large variance, whereas the second one is biased but it has a smaller variance. Here the two criteria of desirability are in conflict. In cases like these an extra criterion is employed, which allows for a trade-off between the two. Mean Squared Error (MSE) is such a meta-criterion. It is the expectation of the square

[36] We briefly touched on this issue in section 1.1.3.
[37] See Newbold (1995) for the mathematical derivations.

difference between the estimator and the population parameter. It can be shown that it is equal to sum of the squared bias and the variance of the estimator. The corresponding rule is to choose an estimator that has the smallest MSE.

We shall see the relevance of the properties of estimators to our discussion in chapters 3 and 4.

2.5 Confidence Intervals

Quite often, however, one is interested in the question as to how confident one should be in the reliability of one's point estimates. Hence there is a method of confidence intervals designed to answer such a question. Confidence interval procedure gives us an interval estimator, rather than a point one, which has a degree of confidence attached to it that the population parameter lies within the interval.

For example, suppose we draw a sample of n observations with mean $\bar{x}$ from a normally distributed population with known standard deviation σ. We would like to find a 95% confidence interval for μ. This confidence interval is given by

$\bar{x} - 1.96\sigma / n^{1/2} < \mu < \bar{x} + 1.96\sigma / n^{1/2}$ Notice that as the number of observations increases, the corresponding confidence interval shortens.

For example, suppose that $X \sim N(\mu, 1)$, and that we have a sample of 36 observations where $\bar{x} = 0.5$. What is the confidence interval for the μ? It is the following:
$0.5 - 1.96/6 < \mu < 0.5 + 1.96/6$, which is $-0.1733 < \mu < 0.8267$.

The usual interpretation of this interval is that if we keep repeating the experiment [i.e., keep drawing random samples from the population], in the limit 95% of the intervals yielded by this procedure will contain the true value of the population mean μ[38]. Hence the *procedure* gives us 95% probability [in the sense of limiting relative frequency] that the intervals contain μ. However, once we have observed a particular

[38] For a representative example, see Newbold (1995):275. For a thorough analysis of the issue of interpreting confidence intervals see Howson and Urbach (2006): section 5.f.2

sample and calculated the particular lower and the upper limits of the associated confidence interval, the frequentist probabilistic interpretation is no longer available to us. This is a manifestation of the general difficulty that the frequentist interpretation of probabilities has with the single-case probabilities. This issue, however, of no consequence to the main issue of this thesis, viz., the problem of model selection.

It is interesting to note [and we shall employ this fact in chapter 4] that the confidence interval estimation procedure is equivalent to the following NP test where c is a constant: H_0: $\mu = c$ vs H_1: $\mu \neq c$. That is, we would reject H_0 at (100% - confidence level %) level of significance just in case c lies outside of the confidence interval. In the example above we would reject the H_0 at 100% - 95% = 5% level of significance if and only if either $c < -0.1733$ or $c > 0.8267$. Here our rejection region is distributed equally to both 'tails' of the distribution [i.e. 2.5% in each tail] in order for the test to be an UMPU. For mathematical details see Lehmann (1986).

2.6 Intermediate Conclusion and Plan

In chapter 1 we introduced the issue of prediction in science. We identified the three ingredients required for this task: data collection, model selection and parameter estimation. We drew distinctions between theoretical and statistical models, and between deterministic and probabilistic models. We stated that this thesis will mainly be concerned with statistical model selection. In this chapter we provided an overview of some widely used model selection methods, viz., Fisherian and Neyman-Pearson. We noted some shortcomings in each of these methods. Chapter three is dedicated to detailed consideration of a relatively new method of model selection that is based on so-called Akaike Information Criterion [AIC]. In chapter four we consider Bayesian statistics in general as it applies to the problem of model selection and relatively novel methodology of model selection based on Bayes Information Criterion [BIC] in particular, and then provide comparison and contrast with the AIC. Chapter five is dedicated to exploring some philosophical consequences of AIC and BIC methods, and in particular to their putative relevance to the debate on scientific realism.

3. The Akaike Information Criterion

3.1 Introduction

> 'So far, when speaking of 'an alternative hypothesis' I have meant some hypothesis genuinely different from the one under test. But in practice Neyman and Pearson do not use 'alternative hypothesis' in such a sense, and this constitutes my second objection to their principle of alternative hypotheses. In practice the alternative hypotheses considered by Neyman and Pearson are nothing but the same hypothesis with different parameter values. Suppose, for example, that the hypothesis under test is that ξ is normal μ_0, σ_0, then the alternatives will be that ξ is normal with different μ, σ (or, in some cases, just with different μ). Thus the alternatives generally considered when the Neyman-Pearson theory is applied are merely trivial variants of the original hypothesis. But this is an intolerably narrow framework. We could (and should) consider a much wider variety of different alternatives. *For example we might consider alternatives which assign a distribution to ξ of a different functional form.*'
>
> Gillies (1973):208, italics added

Let us revisit some of the highlights from chapter 1 relevant to the project of this thesis. There we identified three problems in parametric statistical modelling – coming up with a 'good' sample of data, choosing the model-type and fixing the parameters thus picking a particular model within the model-type. In the rest of this thesis we shall focus on the second issue, viz., the problem of model selection. We will be working on the assumption that we already have a sample of data which has been collected in an acceptable way as discussed in section 1.1.1. The choice of a family of models and estimating the parameters thus picking a particular model within the chosen family quite often goes hand in hand. However, we shall focus on choosing model-types since, even though there are disagreements about how to estimate parameters, the pros and cons of each estimation method are rather well established, whereas there still much more light that needs to be shed on the issue of model selection. We will consider the issue of parameter estimation only when it has a bearing on the issue of model selection.

We think that it would be fair to say that the quote above represents a common perception of the NP framework. In our view, however, the NP framework can be

viewed as providing a method for model selection. Firstly, just because the alternatives generally considered in the NP approach are of the same functional form, it does not imply that they have to be – this is a limitation due to the users of the method, and not of the method itself. Although we are not aware of any actual attempts of NP testing the null and alternative hypotheses of different functional forms, we do not see why in principle this cannot be done. Naturally this would introduce extra mathematical difficulties for, for example, data would be assumed to be arising from different sampling distributions, but this is still a theoretical possibility – cf. Gillies (1973):216. Secondly, even when it is used in the way it commonly is, there are cases when model selection can be said to occur. That is, in the special case of simple null versus composite alternative testing. Admittedly, this is a substantial limitation, although we would reserve the term 'intolerably narrow framework' to the Fisherian methodology. We of course would like to use a broader framework than either Fisherian or Neyman-Pearson for model selection, and indeed this chapter is dedicated to considering one of such frameworks – the Akaike Information Criterion [AIC]. The other framework – that of the Bayes Information Criterion [BIC] – we shall discuss in chapter 4.

Nowadays, there are myriads of methods for model selection. The main reasons why we concentrate our attention on the two methods – the AIC and BIC – are that, firstly, a lot of methods are related to these two, so the methodological and philosophical points that are raised in this thesis by and large carry over, and secondly the AIC and BIC have attracted most attention out of all the other model selection methodologies in recent philosophical literature.

So, the subject of this chapter is the model selection method based on the so-called **Akaike Information Criterion**[39] [AIC].

In order to illustrate the idea of AIC let us come back to one of the examples employed in chapter 1, i.e., to the problem of finding an association between weights and heights of the males within the London Borough of Waltham Forest. Assume that we have collected an admissible sample. Suppose, for ease of introduction, that

[39] Our presentation of AIC is largely based on Burnham and Anderson (2002) and Konishi and Kitagawa (2008)

we have two competing deterministic statistical models – linear and quadratic. By the linear model we mean the infinite set of polynomials of the first degree that have the functional form $y= ax + b$ such that each individual model is an element of this set with the values of parameters *a* and *b* fixed at particular levels. Examples of such models would be linear curves $y = 2x + 3$, $y = 1.6x – 5$, etc. Likewise, the quadratic model is the infinite set of polynomials of the second degree that have the functional form $y= ax^2 + bx + c$ such that each individual model is an element of this set with the values of parameters *a, b* and *c* fixed at particular levels. Notice that in the quadratic model there is no model with the value of parameter *a* set to zero, for such a model would be an element of the linear family. Thus we define our families of models to be *non-nested*. The importance of this point will be discussed in section 4.3.1.

Now, why is it that we are comparing the models *simpliciter*, whereas previously we were often looking at comparing models with the parameters set at particular values as it was the case in Neyman-Pearson methodology in chapter 2? This is because now we focus on comparing the functional forms of models [that is, concentrating on the model selection step] rather than on comparing functional forms of models together with the particular values of parameters.

An obvious way of going about choosing between these two models would be to start with a plot of the data points from the sample in the Cartesian plane such that the heights would be measured along the x-axis and the corresponding weights of the individuals measured along the y-axis. Then, following the most wide-spread approach which urges one to prefer models that reflect the observations as closely as possible, one could find the linear model and the quadratic model that correspondently lie maximally close to the data points. This closeness of fit to data points is conventionally calculated by the sum of the squared vertical distances [hereafter – by the SOS] from each point to the given curve. Unless data point lie on a perfectly straight line, the best fitting member of the quadratic model will provide closer fit to data than the best fitting element of the linear model because of the extra flexibility allowed by having three adjustable parameters [*a, b* and *c*] rather than two [*a* and *b*]. The notion of an adjustable parameter will receive detailed attention further in this chapter. It shall for now suffice to define an adjustable parameter as a

parameter such that every change in its value would pick out a different element within the given model.

Following the reasoning above, a family of polynomials of the third, fourth, and so on degrees would contain elements that provide progressively closer fit to the given data points. This culminates with a perfect fit of a model within the family of polynomials of (*n-1*)th degree, where *n* corresponds to the number of observations that comprise our sample. In this case that best fitting element of this family will go through every data point [as long as there are no data points such that one data point is vertically directly above the other data point], thus having the sum of the squared vertical distances [SOS] at zero. If the closeness of fit is our one and only criterion for choosing a model, we will choose such a polynomial [Forster and Sober (1994):4]. Now, what of our objective of modelling? Recall that we set out in chapter 1 to do modelling for predictive purposes. How predictively successful would we expect the chosen polynomial of (*n-1*)th degree to be? Would we expect the data points from a new sample within the population to lie on or close to the polynomial? Intuitively the answer is 'no', because such a polynomial would have picked up all the idiosyncrasies of the observations making up this particular sample. Even though the sample may have been chosen well – for instance, it may well be representative of the population [which in itself is not a given – recall section 1.1.1], still we would expect the sample to have at least some variation from the population as the whole [again section 1.1.1].

So, why exactly did we get into this trouble with the polynomial of (*n-1*)th degree? One answer is that the corresponding family of polynomial models was too flexible, that is, it contained too many adjustable parameters. Since the closeness of fit increases with more adjustable parameters, it would be natural to think that any given family of models should be penalised for the number of adjustable parameters that the model uses. On the other hand, one would not want the model to have too few adjustable parameters so that the model reflects the given data too remotely. So, there seems to be a trade-off between closeness of fit of a model and the number of adjustable parameters it uses to achieve this fit, with an optimal balance of these two attributes somewhere in-between the two extremes. In fact, this is the notion that one

arrives at through using the AIC methodology, to detailed consideration of which we now turn.

3.2 Components of AIC

In the 1970s the Japanese statistician Hirotugu Akaike derived a formal expression of the idea of the trade-off between the closeness of fit of a model to the data points and the number of adjustable parameters that the models employs to do so[40]. Let us consider his method, which consists of two main components – the maximum likelihood estimation and Kullback-Leibler divergence.

3.2.1 Maximum Likelihood Estimation

Maximum Likelihood Estimation [MLE] is a popular statistical method of estimating parameters given a statistical model form. We delayed consideration of this method to this chapter [the reader will recall that we went through estimation techniques used in classical statistics in the previous chapter] because it naturally aligns with the subject matter of this chapter, i.e., the AIC methodology. The reason for considering the MLE method here is that the Akaike Information Criterion can be viewed as an extension of this method which allows us to not only estimate parameters of the model given the model, but also to choose the model as well.

It is simplest to understand the MLE method by coming back to the Bayes Theorem: $P(H|E) = P(E|H)xP(H)/P(E)$ where 'E' stands for observed evidence [a sample of data, in our case] and 'H' stands for a statistical model with parameters. Recall that we said in the previous chapter that $P(E|H)$ is commonly referred to as the likelihood. That is, $P(E|H)$ is the probability of observing the sample of data at hand given that our statistical model is correct. The MLE method allows us to provide estimates of the parameters of the model on assumption that the model is correct. The methodological prescription in MLE is this: choose the values of parameters in a way that maximises the likelihood. As per usual, in order to comprehend the concept, it is most convenient to look at an example.

[40] Akaike (1973). It is curious to note that this is the same year in which Gillies suggested testing alternative hypothesis with different functional form as per quote at the beginning of this chapter.

Let us again consider the example of throwing a die and noting whether the outcome is an even or an odd number. Suppose that we roll the die 4 times, and that we observed 3 odd and 1 even outcome. Let us define the success rate *p* as the ratio of the number of odd outcomes to the total number of throws. In order to estimate the success rate *p* using MLE we need to maximise P(E|H), where 'E' stands for the observation that 3 out of 4 throws are odd, and 'H' stands for the binomial model with unknown success rate *p*. So, unlike our example in the section dedicated to NP statistics in chapter 2 where we tested null and alternative hypotheses about particular point values of the success rate p, here we would like to estimate this value without any particular ideas as to what it could be.

Recalling the formal expression of Binomial distribution [for details please see chapter 1], the following obtains:
P(3 out of 4 odd | success rate *p*, binomial probability model) = $4!p^3(1\text{-}p)/3!1! =$
$4\ p^3(1\text{-}p)$
We can now conceptualise this expression as a function of the success rate parameter given the observation, say, $L(p|\text{data, functional form of the model})$. This is called a *likelihood function*. Here *p* is variable and data is fixed. Now, in order to find the MLE estimate of *p* we maximise the obtained likelihood function using conventional mathematical techniques, which yields an MLE estimate $p = 0.75$ (recall that $p \in [0,1]$). This means that given the binomial probability model, $p = 0.75$ maximises the probability of observing 3 out of 4 odd numbers. In general, the MLE technique provides parameter estimates that fit the given model as close as possible to the data.

3.2.2 Kullback-Leibler Divergence

Kullback-Leibler divergence [K-L] is the second ingredient necessary to obtain Akaike's result. Kullback and Leibler (1951) derived a measure that aims to calculate the information lost when a given distribution [say, *f*] is approximated by some other distribution [say, *g*]. This information measure [from now on the K-L measure] turned out to be equal to the Shannon's entropy used in information theory [cf. Shannon and Weaver (1949)]. The K-L measure is in the continuous case defined as:

$$I(f, g) = \int f(x) \ln (f(x) / g(x \mid \theta))\, dx \qquad (1)$$

where "ln" stands for the natural logarithm and θ is a vector of adjustable parameters.

The K-L measure is sometimes incorrectly referred to as a 'distance'. It can only heuristically be thought of as a *directed* 'distance' [or divergence] from a model *g* to a model *f*. It is directed because for any model *f* and any model *g* such that if it is not the case that $f \equiv g$ then $I(f, g) \neq I(g, f)$. So it does not satisfy the usual conditions on a distance measure[41]. Hence we shall only refer to the K-L measure as a *divergence* or *information* in the precise sense as provided in the first paragraph of this section. Also, $I(f, g) = 0$ if and only if $f \equiv g$ and for any *f*, *g*: $I(f, g) \in [0, \infty)$.

For illustration, here is an example of using the K-L information for two discrete models. The example is due to Konishi and Kitagawa (2008):33, notation has been modified to fit our usage:

> Assume that two dice have the following probabilities for rolling the numbers one to six:
> $g_a = \{0.2, 0.12, 0.18, 0.12, 0.20, 0.18\}$
> $g_b = \{0.18, 0.12, 0.14, 0.19, 0.22, 0.15\}$
> In this case, which is the fairer die? Since an ideal die has the probabilities $f = \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$, we take this to be true model. When we calculate the K-L information, $I(f, g)$, the die that gives the smaller value must be closer to the ideal fair die. When we deal with discrete random variables, calculating the value of $I(f, g) = \sum_{i=1}^{6} f_i \, ln \frac{f_i}{g_i}$ we obtain $I(f, g_a) = 0.023$ and $I(f, g_b) = 0.020$. Thus in terms of K-L information, it must be concluded that die g_b is the fairer of the two.

Now, if we interpret distribution *f* as the truth [or the actual data generating distribution, or some such like notion – we will leave a more careful consideration of this notion until the next chapter] and *g* as a model which is used to approximate *f*, and in addition assume that the truth is fixed, then under some general conditions [cf. Burnham and Anderson (2002) for a fully rigorous mathematical derivation] Akaike (1973) established that a relative divergence from *g* to *f* can be estimated by the

[41] A function is a distance measure if for any three vectors **l**, **m** and **n**, it assigns positive real numbers *r* subject to the following conditions:
$r(\mathbf{l}, \mathbf{l}) = 0$; $r(\mathbf{l}, \mathbf{m}) = r(\mathbf{m}, \mathbf{l})$ [symmetry]; $r(\mathbf{l}, \mathbf{m}) \leq r(\mathbf{l}, \mathbf{n}) + r(\mathbf{n}, \mathbf{m})$ [triangular inequality]

maximised log-likelihood function $\ln(L(\hat{\theta} \mid data, g_i))$ for each model g_i, [i = {1,…,r}] from the set of *r* models in the choice set. However, Akaike found that such a maximum likelihood estimator is asymptotically positively biased [cf. section 2.4.1], and that in large samples the bias is approximately equal to *K* – that is, the number of adjustable/estimable parameters in the model.

Then by multiplying $\ln(L(\hat{\theta} \mid data, g_i)) - K$ by (-2)[42] he defined the **Akaike Information Criterion:**

(AIC) $$-2\ln(L(\hat{\theta} \mid data, g_i)) + 2K.$$

Model selection using AIC proceeds thus. First of all, a set of competing models is compiled on the basis of the background information, theoretical results/ideas in the field or research, previous research, etc. Hereafter we refer to such a set as the *choice set*. Then the data are considered.[43] Then in each model the adjustable parameters are set at their maximum likelihood levels so that an element of each model is obtained such that it provides the closest fit to the given data. Then the AIC scores are calculated for each of these closest fitting elements. In a sense the closest fitting elements of each model represent their respective models. The model which has the maximum likelihood element with the smallest AIC score is chosen.

Since we do not know the "full reality" *f*, only the differences in the AIC scores between the models in the choice set are interpretable, and not the absolute values. When considering the differences, the constants *C* cancel out, so Akaike scores are on an interval scale lacking a true zero but preserving the relative distances [i.e., ratios of distances], whereas the K-L information itself is measured on a ratio scale with a true zero. So, AIC gives us an expected directed K-L distance from the given model to the unknown full reality *relative to* models in the choice set, and no others. This means that by using AIC we do not have epistemic access to the directed distance from models to the truth in the absolute sense. This highlights the importance of picking the models for the choice set with the utmost care and

[42] Multiplication by (– 2) was done for "historical reasons". For instance, under certain assumptions, $-2\ln(ML_1/ML_2)$ is asymptotically χ^2 distributed.

[43] The sample could have been gathered prior to the compilation of the choice set or afterwards. This is irrelevant so long as the data were not used in creation of the choice set. For the discussion of this point please see section 3.4.1.4 below.

consideration for the background information and available experience and knowledge in the field at hand. This is what we meant when we drew a distinction between theoretical and statistical modelling [cf. section 1.1] and said that the distinction is not sharp because we are concerned with statistical modelling which has theoretical elements in it. These theoretical elements play their part when one picks the families of models to compile the choice set.

Due to the meaningfulness only of AIC differences, it is recommended to calculate the **AIC differences**, $\Delta_i = AIC_i - AIC_{min}$, for each model in the choice set. These are estimates of the expected K-L differences from $g_i(x \mid \theta)$ to f relative to the model, which has the smallest AIC score [denotes by AIC_{min}]. The best estimated model has $\Delta_i = \Delta_{min} = 0$. There is always at least one best AIC estimated model within the choice set. The Δ_i values allow for ranking of models within the choice set. The (naïve) methodological rule is: choose the model with $\Delta_i = \Delta_{min} = 0$. A refinement of this rule is considered further.

In order to work out the relative strength of evidence for each model, the likelihood $L(g_i \mid x)$ of model g_i, given data x, is defined in the literature to be proportional to exp (- 0.5 Δ_i)) Then, for ease of interpretation, all $L(g_i \mid x)$ for each model in the choice set are normalised to yield so-called "Akaike weights", w_i, which all add up to 1.

$$w_i = \exp (- 0.5 \Delta_i) / \Sigma^{n}_{i=1} \exp (- 0.5 \Delta_i)$$

Burnham and Anderson (2002) refer to these weights as "model probabilities" or "the weight of evidence in favour of model *i*". Akaike weights ratios are equal to relative likelihood ratios [i.e., for a pair of models *i* and *j*, $L(g_i \mid x) / L(g_j \mid x) = w_i / w_j$], which are in the AIC literature taken to '…represent the evidence about fitted models as to which is better in a K-L information sense.' [Burnham and Anderson (2002):78]

Let us consider an example of actually using the AIC method for model selection. Naturally, it would be great to develop further one of our earlier examples, say, the one from chapter 1 on relating weights and heights of male residents of the London

Borough of Waltham Forest. Unfortunately we do not have any data for our mock example, so we shall have to use a different example which is structurally similar to our weights/heights e.g. In any case we can use the weights/heights example to introduce the actual example we shall use.

So, starting with the simplest case, suppose that we come up with a probabilistic statistical model for our weights/heights example and suppose that this model is of a linear regression type. That is, weights [the response variable Y] and heights [the regressor variable X] are linearly related thus: $Y = aX + b + \varepsilon$, where a is the gradient of the line, b is the intercept with the y-axis and ε is the residual error term which accounts for the deviations of data from our linear model. It is commonly assumed that the deviations from the line are probabilistically independent from one another [cf. section 1.3], and that ε is normally distributed with zero mean and a constant standard deviation σ, where σ is estimated from data. We shall go along with this assumption. Suppose that we wish to use the least squares method of linear regression. That is, we find the element of our linear model in such a way that the sum of the vertical distances from this element to the data points [SOS] is the smallest among all the elements of our model. The least squares linear regression is in fact a special case of general maximum likelihood estimation. We can thus obtain the AIC scores with the output of standard regression packages using this formula:

$$\mathbf{AIC} = n \ln(\hat{\sigma}^2) + 2K$$

where K is the number of estimated regression parameters including σ^2 [in our case there are three adjustable parameters – a, b and σ]; $\hat{\sigma}^2$ [the estimated variance] is equal to its maximum likelihood estimator ($\Sigma\hat{\varepsilon}_i^2/n$).

However, our mock example of weights/heights is very artificial. Although it seems reasonable to think that weights increase with heights, trying to predict weights with heights seems insufficient. For instance, we may also wish to include the dietary preferences [on the thinking that those with preference to foodstuffs that contain more energy would be heavier], the weight of the mother and the weight of the father reflecting the idea that our males' weight could possibly be related to that of their parents. There can be many other variables we may wish to consider. Now instead of simple linear regression we have a multiple regression case: $Y = aX_1 + bX_2 + cX_3 +$

$dX_4 + \varepsilon$, where X_1 stands for the height, X_2 – for the preference for particular type of food [perhaps as measured by the average amount of kilocalories such food contains], X_3 and X_4 – for the weights of the mother and the father respectively. Now, which variables are relevant for predicting the value of Y? In the absence of further knowledge, it seems that we have $2^4 - 1 = 15$ possible models to choose from, assuming that at least one variable is relevant.

So, here is a much used example which is commonly referred to as Hald's Cement Hardening Data – several references are cited by Burnham and Anderson (2002):99-103. This example is structurally the same as our multiple regression case above. The table below represents heat evolved during the hardening of 13 samples of Portland cement and four variables that may be related to it – the tables are from Ghosh and Samanta (2001).

Cement hardening data with four regressor variables x_1, x_2, x_3 and x_4 and a response variable y

x_1	x_2	x_3	x_4	Y
7	26	6	60	78.6
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

where the regressor variables (in percentage of the weight) are: x_1 = calcium aluminate (3CaO.Al2O3), x_2 = tricalcium silicate (3CaO.SiO2), x_3 = tetracalcium alumina ferrite (4CaO.Al2O3.Fe2O3) and x_4 = dicalcium silicate (2CaO.SiO2); the response variable is y = total calories given off during hardening per gram of cement after 180 days.

Our purpose is to select a model for predicting the evolution of heat in Portland cement on the basis of its chemical composition. We assume no detailed knowledge of physics or chemistry, and so engage in probabilistic statistical model selection [as opposed to theoretic]. Thus we put the 15 possible models in our choice set [of course we could also consider models with quadratic, cubic etc. terms, but that would be unnecessary for our purposes] and calculate Akaike differences AIC (Δ). That is, the model with AIC (Δ) = 0 is deemed AIC-best. In the table below the first column indicates the type of model by showing which variables are included in each model. For example, the model in the first row has only x_1 and x_2 in it and thus it has four adjustable parameters [$K = 4$] – the two parameters that are multiplied by the variables, the intercept with the y-axis and the variance. Another point to note is that below there is a column for AIC_c, which is a version of AIC used when the number of data points are small [remember that AIC is an *asymptotically* unbiased estimator of relative expected K-L divergence] relative to the number of adjustable parameters used in the 'maximal' model in the choice set – viz., the model which has the highest number of adjustable parameters of all models in the choice set. Sugiura (1978) and Hurvich and Tsai (1989) found that when the ratio of the sample size to the number of adjustable parameters in the maximal model is small [some consider that this is the case when the ratio is below 40 – cf. Burnham and Anderson (2002):66] there is a small sample bias which requires a [second order] correction. An intuitive explanation for this bias is that when the ratio is small there are more adjustable parameters than can be justified with such limited data. So, AIC_c penalises models that use extra adjustable parameters relative to other models in the choice set disproportionately more than does the AIC. Our sample consists of only $n = 13$ observations, and the maximal model in the choice set has six adjustable parameters, so it is more appropriate for us to use AIC_c for model selection. All the methodological points about AIC carry over pretty much verbatim to AIC_c.

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$$

In the table below the model which has only variables x_1 and x_2 in it is AIC_c-best. The maximum likelihood (ML) element [i.e., the element that fits the data most closely] of this model is:

$y = 52.6 + 1.468x_1 + 0.662x_2$, and $\hat{\sigma} = 2.11$ [Burnham and Anderson (2002):103]

Also notice that there are some models that are not that far from the AIC_c-best one.

In particular, models in rows two, three, four and five all have the AIC_c differences below 4. According to the 'rule of thumb' that is used in AIC methodology [which applies equally to both AIC and AIC_c], these models also have some support. The rule of thumb is that models that are within 2 units of the AIC-optimal model have substantial support, those that are between 4 and 7 units away from the AIC-optimal model have considerably less support, and those that are more than 10 units away have virtually no support at all [Burnham and Anderson (2004):271].

Model	*K*	AIC (Δ)	AICc (Δ)
12	4	0.45	0
124	5	0	3.13
123	5	0.04	3.16
14	4	3.77	3.32
134	5	0.75	3.88
234	5	5.6	8.73
1234	6	1.97	10.52
34	4	14.88	14.43
23	4	26.06	25.62
4	3	33.88	31.1
2	3	34.2	31.42
24	4	35.66	35.21
1	3	38.55	35.77
13	4	40.14	39.7
3	3	44.09	41.31

Ghosh and Samanta (2001):1143

3.3 Some Features and Properties of AIC

Here are some properties of the AIC, some of which may seem self-evident.

- AIC differences between models based on different sets of data cannot be compared.
- The order of computation of AIC scores is irrelevant.
- Models that are not in the choice set are out of the consideration.

Probably the most important feature of the AIC methodology is its use of K-L divergence. However, there are several alternative measures of discrepancy between distributions [cf. Konishi and Kitagawa (2008):31]. Is there a justification for using the K-L divergence rather than any other measure of divergence or distance between distributions?

Burnham and Anderson (2002) assert that 'the relative K-L distance is the link between information theory and the log-likelihood function that is a critical element in AIC model selection.' [Burnham and Anderson (2002):87] 'The K-L distance between models is a *fundamental quantity* in science and information theory … and is the logical basis for model selection in conjunction with likelihood inference.' [Burnham and Anderson (2002):54]

Burnham and Anderson's (2002) argument for the use of the K-L discrepancy in model selection rather than any other measure is in its essence nothing over and above an argument by analogy – roughly, success in some fields implies success other fields. Their argument for the special status of K-L discrepancy is two-fold. Firstly, this quantity has its natural place in information theory [e.g., Shannon information entropy], which they consider to be a fundamental advance in 20th century science. Secondly, entropy is of fundamental importance in statistical mechanics. So, the former seems to assert that K-L actually arises from the information theory and the latter is an argument by analogy. They site as important Bolzmann's theorem connecting entropy to negative logarithm of probability.

With regards to the latter argument we agree with Jaynes (1957):621:

> 'The mere fact that the same mathematical expression $-\sum p_i \log p_i$ occurs both in statistical mechanics and in information theory does not in itself establish any connection between these fields. This can be done only by finding new viewpoints from which thermodynamic entropy and information-theory entropy appear as the same *concept*.'

On the matter of asymmetry of K-L discrepancy measure Burnham and Anderson [arguably the authors of the most definitive and up-to-date work on the subject of AIC – i.e., Burnham and Anderson (2002)] say only the following: '$\ldots I(f, g) \neq I(g,$

f); *nor should they be equal, because the roles of truth and model are not interchangeable*.' [Burnham and Anderson (2002):56] It is hard to say what to make of this remark. One may try to interpret it in a way that approximating a model by truth is not a sensible thing to do because it is the approximation the other way around that interests us, hence, calculating the distance makes sense in one direction only. However, granting to the truth the special status, it still does not constitute a reason as to why a distance *to* it should be any different from the distance *from* it. This asymmetry seems to a natural interpretation in thermodynamics as an increase in entropy, and it represents the arrow of time: to go back to the previous state requires more energy than to go from it. However, in our context of model selection there is no obvious reason of this sort for the asymmetry of our divergence. So much for the argument from analogy with thermodynamics!

We find that the most convincing argument for using the K-L divergence rather than any other is that the K-L divergence lends itself easily [and some may say naturally] to approximation by the ML technique, which is well-established within modern statistics.[44] Still, the lack of symmetry is worrisome and should be born in mind as a shortcoming.

[44] According to Akaike himself, the connection occurred to him in March 1971 when he was standing on the train from his home to the institute where he worked at the time [Findley and Parzen (1995):111].

3.4 Philosophical Issues with AIC

Recent philosophical literature contains/identifies several issues with the AIC methodology. Issues considered in this section are to do with adequacy of AIC as a model selection methodology, its use and limits. Some external applications of AIC in broader philosophy of science context are dealt with in chapter 5.

3.4.1 The Subfamily Problem

3.4.1.1 Statement of the Problem

The problem that is identified in this section is related to the issue as to where we get the models from to compile the choice set.

The subfamily problem [identified by Forster and Sober (1994)] can be explained in the following manner. Suppose that a model that we pick for the choice set is an (*n*-1)th degree polynomial such that it contains an element which perfectly fits the *n* data points that we have [i.e., SOS = 0]. Moreover, this particular polynomial is the only element of the given model [that is, the model constitutes a singleton set]. This 'model' and its element will be chosen by an information criterion[45] as optimal whatever the alternatives since it has no adjustable parameters [$K = 0$] and it fits the data perfectly!

3.4.1.2 The Forster-Sober Solution

Forster and Sober's (1994) apparent solution of this problem is based on what they call the Error Theorem and the distinction that they draw between statistically unbiased and epistemically unbiased estimation.

The Error Theorem: Error[Estimated(A(*F*))] = Residual Fitting Error + Common Error + Sub-family Error.[46] Here A(*F*) denotes the predictive accuracy of the family of curves *F*.

[45] All information criteria suffer from this problem – cf. Forster and Sober (1994):18, fn 27.
[46] Forster and Sober (1994):19

An estimator is *statistically* unbiased if and only if its expectation is equal to the actual value of the parameter that it is used to estimate. The idea of *epistemic* bias is best described by means of an example. Suppose that we have a statistically unbiased estimator. Let us increase its variance arbitrarily in such a way that the estimator's mean value is unaffected. We still have a statistically unbiased estimator. But, argue Sober and Forster, it is epistemically biased since there is at least one other unbiased estimator, which has a smaller variance than the one at hand. Forster and Sober argue that the *ad hoc* application of Akaike's Theorem to the singleton models as described above is statistically unbiased but epistemically biased, and that this is implied by the Error Theorem. Let us see how.

The Common Error is the same for all the models, so it cancels out. The Residual Fitting Error is both statistically and epistemically unbiased. However, the Sub-family Error is statistically unbiased but sometimes epistemically biased.

Forster and Sober illustrate how this epistemic bias arises in the following way.

> Suppose we have a very large data set that exhibits strong linearity. We wish to estimate the predictive accuracies of L(LIN) and L(POLY-n), where POLY-n is the family of n-degree polynomials with n parameters free, and $L(F)$ is obtained by using the data to single out the best fitting curve in family F. We may apply Akaike's Theorem to (LIN) and (POLY-n) *directly*, or we can apply it to the singleton families containing just L(LIN) and L(POLY-n), respectively. The surprising fact – that the *ad hoc* Akaike's estimate for L(POLY-n) and L(LIN) will always favour L(POLY-n), because it is always closer to the data. In sum, both the direct and the *ad hoc* method of accuracy estimation are statistically unbiased (as required by Akaike's Theorem), but the *ad hoc* application of Akaike's method yields an estimate that we *know* is too high. The *ad hoc* application yields an estimate that is *epistemically biased*. [Footnote 31: Although the estimate is known to be too high, given the data at hand, the Akaike estimate of the predictive accuracy of that same singleton family relative to *other* data sets generated by the true 'curve' will be too low. On average, of course, the estimate will be centred on the true value.]
>
> Forster and Sober (1994):21

Forster and Sober say that the Error Theorem is in fact a 'meta-theorem' – it is a theorem about the 'meaning' of Akaike's Theorem.[47] They state that this

[47] *Ibid.*

result is closely related to the one in Sakamoto *et al*. (1986):77. So, let us consider it[48].

-½AIC(*K*) = (mean expected log likelihood) + (common error) + (individual error)

Let us analyse this result. The common error does not depend on the number of adjustable parameters, *K*, in a given model, so *K* does not have a bearing on the model selection. The individual error is a sum of two expressions. Let individual error = (C + D)[49]. C's variance is equal to *K*, and C increases as *K* increases. However, D decreases with increase in *K*. Now, due to the subfamily problem, *K* = 0. Hence the individual error does not have a bearing on the model selection either. But the mean expected log likelihood of *L*(POLY-*n*) is higher than that of *L*(LIN). Thus, the AIC(*L*(POLY-*n*)) is smaller than AIC(*L*(LIN)). So the AIC methodological rule prescribes the choice of *L*(POLY-*n*). Therefore, we conclude that using Sakamoto's result does not solve the subfamily problem. Hence, it appears to be the case that the closely related Error Theorem does not solve it either.[50] There is a reason to think that the *ad hoc* application of Akaike's Theorem is epistemically biased, viz., our perception that *L*(POLY-*n*) picks up too many errors by fitting the data too closely. But we argue that the epistemic bias is not implied by the Error Theorem. So, we are seemingly back where we started from – the subfamily problem. We thus conclude that Forster and Sober (1994) do not succeed in resolving the subfamily problem by using their Error Theorem and the distinction between the statistical and epistemic bias.

3.4.1.3 Replies from Kukla and Kieseppä

Kukla (1995) starts off with noting that: '(1) …there are infinitely many equally good candidate-curves relative to any given set of data, and (2) … these best candidates include curves with indefinitely many bumps.' [Kukla (1995):248] Presumably by 'equally good candidate-curves' Kukla means models that fit the data equally well, but differ in their predictions of future data.

[48] This part follows Sakamoto *et al*. (1986): 76-81.
[49] For the full mathematical rigor cf. *Ibid.*
[50] Indeed, Kieseppä [(1997): 40] aptly remarks on this argument from Error Theorem: 'This is a clever argument, but the unrigorous way in which it has been presented makes it very difficult to evaluate whether it really solves the subfamily problem.'

So, the first problem is: just SOS fitting with (*n*-1)th degree polynomial allows for any prediction whatsoever; the second problem: linear relationships would never be used contrary to common scientific practice. Kukla states that Forster and Sober (1994) ignore the first problem and concentrate on the second.

Kukla raises the following issue. Take families of models which contain as their elements polynomials, say, of (*n*-1)th degree such that they have (*n*-1) adjustable parameters [i.e., one of the parameters is fixed] and the best fitting element in each such family has the SOS equal to that of the best fitting element of the family that contains polynomials of (*n*-2)nd degree with (*n*–1) adjustable parameters. AIC would give these two models an equal score. Importantly, both models have the same Sub-family Error [as per Forster and Sober – cf. section 3.4.1.2], but the elements in the former have an arbitrary number of bumps. So, seemingly we do not have an epistemic criterion for showing that a polynomial of a degree lower than another and the same SOS is epistemically preferable/predictively more accurate.

As an example, consider linear and quadratic functions. Pick a quadratic function with one fixed parameter [thus the number of adjustable parameters that are left is two] and adjust the remaining adjustable parameters in such a way that the expression has the same SOS as the linear function. There are in fact infinitely many expressions of this sort [we can repeat the procedure with polynomials of higher and higher order].

Forster (1995a) says that there is nothing wrong with having infinitely many curves with the same predictive accuracy. The problem arises when the criteria consider curves predictively equally accurate whereas they are in fact not [In footnote 2 page 349 he says that a bumpier curve could be closer to truth if the truth were bumpy, but on average would not be.]: '…Kukla appeals to the intuition that very bumpy curves are not expected to have equal predictive accuracy…I concede that Kukla's intuition is correct.' [p. 349] Forster[51] addresses the second problem raised by Kukla by

[51] In what follows Best(PAR) stands for the actual truth; L(PAR) – best fitting member of the model of all parabolas; L(LIN) – best fitting element of the model of all linear curves; Qi – a model of 2nd

devising a geometric example and showing that it can be interpreted interpretation within the Akaikean framework in the case of curve fitting, on

> the key assumption … that the location of L(PAR) is governed by a Gaussian (i.e. Normal) distribution centred at Best(PAR) with a variance inversely proportional to the number of data. As a result L(PAR) will stray less from Best(PAR) as data accumulate. … Note that this 'normality' assumption does not require that the noise in the data itself is Gaussian. [Footnote 8: 'Kukla's presentation is potentially misleading in that he talks as if the sum of square deviations (SOS) is always the appropriate measure of fit, but this is only the case for Gaussian errors. AIC uses the general measure of log-likelihood to measure fit, as we made clear in Forster and Sober [1994].'] The assumption is about the *effect* of noise in parameter space… The significance of this 'normality' assumption is that it licenses a geometrical interpretation of hypothesis space.
>
> Forster(1995b):353-354

Forster's interpretation of the geometric example shows that a randomly selected L(Qi) will do worse [in the sense of being on average less predictively accurate] than L(LIN). Going through the mathematical details of his Theorem, Forster states that '[a] remarkable feature of this result is that the average advantage of LIN over Qi does not depend on the amount that curves in PAR are capable of performing better than anything in LIN.' [Forster(1995b):356]

Forster proposes a modification of the AIC in order to correct for the problem introduced by Kukla's way of choosing families of models for the choice set. According to him, the AIC score for the polynomial should be increased by $\Delta K/n$, where ΔK is the increase in the number of adjustable parameters and N is the number of data points. In his subsequent papers on the Akaike methodology [and information criteria in general[, however, Forster does not include the proposed correction of AIC[52]. This seems to indicate that correcting AIC measure by quantity $\Delta K/n$ has to be used when among the models that one considers at choice set step are those of the Qi type. Moreover, in order for the reply to Kukla to work, one has to *randomly* select a model Qi among the models of its type to be considered the choice set.

---

degree polynomials with one fixed parameter; L(Qi) – the best-fitting element of Qi, which is equal in simplicity and fit to L(LIN).

[52] Nor is it generally used by statisticians.

Even if we accept Forster's answer as a partial solution to Kukla's challenge, Kieseppä (1997):39-40 poses a problem to which Forster's geometrical construction has no answer. Kieseppä considers a situation where one happens to include a model of Qi type in the family of models for consideration *prior* to observing data. Then Forster's solution does not apply, but arbitrariness in the choice of models for the choice set remains. Would Forster call this arrangement *ad hoc*? Kieseppä states that to choose the hierarchy, we seem to require the knowledge about what good scientific hypotheses look like, which does not stem from mathematical theorising.

3.4.1.4 Our Own Dissolution of the Problem

Interestingly, we have not come across the subfamily problem anywhere in the extensive literature aimed at statisticians, who are interested in foundational issues as well as in application of the statistical techniques. Perhaps this is due to this issue not being seen as a serious problem. We think that it is not a serious problem, although as we argue in sections 3.4.1.2 and consider in section 3.4.1.3 it has evaded proper resolution hitherto. In fact we go as far to argue that this is not a problem for model selection – we dissolve it.

Firstly, suppose that we put in the choice set a model which contains all linear functions as its elements and a model which only contains a single element, e.g., $y = 2x^2 + 3x - 5$. Why would we want to include the latter model? There are two potential reasons – either we have had a preliminary analysis of data and found out that this is the best fitting parabola out of the model of all parabolas or we have good reasons to think that this is a good model on the basis of currently accepted theories, our experience in the field, etc. If we do it for the first reason, then we suggest that a 'counterfactual' move could be made in order to stop the subfamily problem from appearing. That is, one should not check the data first – the models should be chosen for the choice set on the basis of the background knowledge that we have prior to observing/considering the data set that we are using for model selection. And if one is already quite familiar with the data, one should 'forget' that one is familiar with it. Arguably, even if we already have collected our data sample, we should not attempt to reflect the sample in the hypotheses in our choice set. Even if we have strong familiarity with the data, we should 'delete' it or 'forget it' when considering which

models to include in our choice set. For instance, we are quite surprised that Howson and Urbach (2006) do not adopt this stance on the subfamily problem rather than calling it '…this rather devastating objection.' [Howson and Urbach (2006):294] On the contrary, this argument is analogous to, and our stance is very much consistent with the one adopted by them with regards to the Old Evidence Problem in Bayesian Confirmation theory[53]. Here is a brief rendition of the Old Evidence Problem [for more details see Howson and Urbach (2006):297-301; for an overview of other attempts to solve the old evidence problem cf. Earman (1992):chapter 5]. It is commonly thought in the philosophy of science that if you build a hypothesis to entail known data, that hypothesis cannot draw any support for itself from that data. Only new data can confer confirmation onto a hypothesis. Sometimes, however, new theories are not purpose-built to fit old data, but once they are developed *independently* of the already known data, on occasion it is post factum found that they do fit old data. It is commonly thought that in such a case the old data supports the new hypothesis. However, in Bayesian confirmation theory [which, very briefly, is the idea that data E confer evidential support onto a hypothesis H when the posterior probability of the hypothesis H in the light of data E is greater than its prior probability] the probability of data that has already been observed [called it 'E_{old}'] is $P(E_{old})=1$ and also the likelihood of E_{old} is $P(E_{old}|H)=1$. Hence using the Bayes Theorem, $P(H|E_{old})=P(H)$, so the old evidence does not confirm the hypothesis, contrary to our intuition. Howson and Urbach propose a counterfactual move to remove the evidence implied by the hypothesis from the background information, on which all the propositions in Bayesian theory are conditional. Then the old evidence can provide support to the hypothesis H.

A closely related idea is the use-novelty account of support of hypotheses by evidence [cf. Worrall (2002)].

> 'A fact will be considered novel with respect to a given hypothesis if it did not belong to the problem-situation which governed the constitution of the hypothesis.'
> Zahar (1973):103

[53] We think it would be fair to say that Howson and Urbach (2006) do not consider the Old Evidence Problem to be a problem at all, and find it incredible to see that so much effort has been expended on trying to resolve it. Again, we find ourselves in an analogous position with regards to the Subfamily Problem.

Under this account only novel facts in this sense provide support to hypothesis. So on this idea using the singleton hypothesis $y = 2x^2 + 3x - 5$ would be fine as long as the given sample of data has not been used in order to construct this hypothesis [irrespective of the period of time in which such a sample was collected]. If this element of the parabolic model then would provide a perfect fit, that would be absolutely fine.

Secondly, we have good prior reasons for choosing the particular values for the parameters only if we already have a good idea as for the functional form of the relationship between variables. In other words, there is hardly any model uncertainty. Since the problem of model selection is essentially the problem of model uncertainty, there is no place for model selection and hence for model selection criteria's use in such a case. So, this defeats the very purpose of model selection. Employing this procedure is akin to doing the following. Instead of carefully selecting a small number of competing hypotheses on the grounds of our background knowledge and theoretical research and then observing the data, we are now going to have a thorough trawl through the data and find the best fitting model of an arbitrary dimension. Then we shall form a model consisting of this singleton model and retro check whether it obtains the highest AIC score among any other possible models. And then – low and behold – we will find out that it does! The question then arises as to the purpose of such an exercise – we know in advance that such a procedure would give us the top AIC score, whatever the data we are going to observe. Of course this emphasises the logical point of the subfamily problem, but in the process it defeats the very purpose of model selection, which is to choose the optimal mathematical structure of a model for predictive purposes because there is uncertainty as to what this structure should be, and some prima facie viable alternatives available.

Finally, even if one might find the 'counterfactual' move unappealing, we think that for the purposes of model selection in the case when the 'artificial' fixing of singleton hypotheses within the choice set takes place, the sense in which K is the number of *adjustable* parameters should be that "capable of being adjusted at some point in time", rather than just "free to be adjusted now". Hence, a model containing

$y = 2x^2 + 3x - 5$ as its sole element would still have three adjustable parameters in the relevant sense. This is because it is an element of the model with three adjustable parameters. From this point of view, in section 3.4.1.3 L(Qi) would have more adjustable parameters than L(LIN), since we deliberately fixed (importantly, at a non-zero value) an adjustable parameter in L(Qi). Problem dissolved.

3.4.2 The Problem of Language Variance

3.4.2.1 Grue Problem

De Vito (1997):392 makes a two-fold claim: 'The problem with using Akaike's theorem for hypothesis choice is that the number of parameters associated with a given hypothesis is a matter of *convention*. In addition, for any hypothesis there is no a priori way to generate the *right* family of curves to which the hypothesis belongs.'

De Vito demonstrates the former claim by applying the information criteria (he focuses on AIC, but the argument, if correct, would also apply to BIC and other information criteria) to Goodman's New Riddle of Induction.

Here is the argument. Suppose that we hypothesise as to the colour of emeralds over time. Let us define a predicate Grue such that 'object *x* is grue at time *t* if and only if *x* is green at time *t* and $t < 2100$, or *x* is blue at time *t* and $t \geq 2100$.' [Forster (1999): 92] Hence, we have two hypotheses regarding the properties of emeralds:

Green Hypothesis: 'All emeralds are green (at all times).'
Grue hypothesis: 'All emeralds are grue (at all times).'

These hypotheses fit the current data equally well, but De Vito argues the Grue hypothesis contains one adjustable parameter [viz., *t*] whereas Green hypothesis has none, so AIC would lead us to favour Green hypothesis. Now, define predicate Bleen such that 'object *x* is bleen at time *t* if and only if *x* is blue at time *t* and $t < 2100$, or *x* is green at time *t* and $t \geq 2100$'. [Forster (1999):94] Note that in the language in which the predicates Grue and Bleen are taken to be ordinary, an '…object *x* is green at time *t* if and only if *x* is grue at time *t* and $t < 2100$, or *x* is bleen at time *t* and $t \geq$

2100.' [Forster (1999):94] Hence, Green hypothesis becomes: 'All emeralds *x* are such that, if $t < 2100$ then *x* is grue at time *t* and if $t \geq 2100$ then x is bleen at time *t*.' [Forster (1999):94] whereas the Grue hypothesis in this language is still the same: 'All emeralds are grue (at all times).' Now, in this new language both hypotheses fit the data equally well, but now the Green hypothesis has more parameters than the Grue one. So, by application of AIC, in this language one should favour the Grue hypothesis. The number of adjustable parameters that models have depends on the particular conceptualisation of the world. Hence, information criteria suffer from language variance.

Forster (1999) replies to this argument by agreeing with De Vito that the application of AIC does not solve Goodman's New Riddle of Induction and that this problem is a curve-fitting one. Forster argues, however, that De Vito draws from this an incorrect conclusion. The correct conclusion is that AIC does not apply to all curve-fitting problems.

Forster argues that De Vito misconstrues the notion of adjustable parameter. In fact, neither the Green nor the Grue hypothesis contain any adjustable parameters in the sense that this notion is used in Akaike's methodology. A model contains adjustable parameters just in case a change in these parameters will pick out a different element in the model. In the case discussed by De Vito, the competing models are singleton sets [containing exactly one element respectively], hence all the parameters are adjusted. Exactly the same applies to the hypotheses when they are described in the 'Grue' language. So, in either case AIC is unable to distinguish between the hypotheses. Another problematic aspect of posing the problem the way that De Vito does is that AIC applies only to probabilistic hypotheses: 'The concept of fit in Akaike's theorem is derived from the Kullback-Leibler discrepancy, which requires that the competing hypotheses are probabilistic (so that the likelihoods are well-defined.)' [Forster (1999):93] We could turn Green and Grue hypotheses into probabilistic ones by assuming that the observation errors are probabilistic. Even if we do so, AIC will not give us any reason to prefer one hypothesis over the other, which is contrary to our intuitions that Green hypothesis should be preferable to Grue hypothesis.

Forster modifies De Vito's example in such a way that Grue hypothesis does contain an adjustable parameter –instead of fixing time parameter at value of 2100, it is now $t = \theta$. Now, does AIC tell us to pick Green hypothesis rather than Grue? No, it does not, because '…Akaike's notion of simplicity aims to quantify the sampling error in the parameter estimates. But in this example, there is still no *sampling* error in the estimation of the grue parameter θ. …[T]he grue model is unidentifiable in the sense that there is no unique value of θ that maximizes the fit with the seen data. There is no over fitting or under fitting in the relevant sense.' [Forster (1999):96] Intuitively, AIC helps us when we have to predict future data by considering the observed data and stipulating that the future data comes from the same distribution. In this case, however, no such assumption can be made – we have no idea how (if at all) the distribution of θ is connected to the distribution of the observed data.

3.4.2.2 Reparametrisation under Transformation

De Vito poses a more general argument than that considered in the previous sub-section. He gives an example in virtue of which under a certain transformation of the coordinate system a family of parabolic functions (PAR) becomes a linear one (LIN), and a linear family becomes parabolic. On the assumption that the SOS of the perspective best fitting elements of both is the same, De Vito argues that the AIC will recommend different curve in each situation. De Vito concludes that in virtue of his results a realist solution to the curve fitting problem is not warranted since the closeness to truth cannot be relative to a particular conceptualisation of the world.

Forster replies to this charge as well. 'The main problem is that transformations do not map a single member of PAR into a unique member of PAR', so there is no sense in which the transformed families are *equivalent* representations of the old families.' [Forster (1999):95] The argument is based on the assumption, which is used in the derivation of Akaike's Theorem. That is, if F is a subfamily of G that F cannot be more complex than G and this subset relation is preserved under any one-to-one transformation. So, if F is a subset of G then F' is a subset of G' and so is less than or equally complex.

As a part of his argumentation, Forster rather informally goes through the first part of a proof of Akaike's Theorem. An intermediate step in the proof is an estimation of the discrepancy between the curve that fits best the observed data and the true curve [$\Delta(\theta)$] by Taylor-expanding this discrepancy around θ^* in terms of $\Delta(\theta^*)$, [where θ^* is presumably the best fitting element of a given model (Forster (1999) does not define what he takes θ^* to be)] and showing that Taylor expansion is language invariant. Having gone through the theorem, Forster concludes that '…[*K*] is not simply the number of adjustable parameters, but the number of parameters *that contribute to the expected discrepancy in a certain way*. Given the fact that Taylor expansion is language invariant, and expected values are language invariant, there is no way in which this number can change by any redescription of the families of the curves. … It is convenient to describe *k* as equal to the number of adjustable parameters only because the equality holds in most cases. … [L]anguage invariance is built in at the very beginning.'' [Forster (1999):100]

Kieseppä (2001b) comments that one has to fix the representation in which one makes decisions. In Bayesian context, if one has not fixed a particular representation then one cannot use the difference in the visual simplicity of curve in order to fix the priors – lower for more 'complicated' curves and higher for 'simpler' ones, because different polynomials can have identical mathematical properties under a transformation. [Kieseppä (2001b):784]

4. Bayesian Statistics and the Bayes Information Criterion Methodology

In the previous chapter we considered the AIC model selection methodology and defended it against various objections brought to bear in the literature. The purpose of this chapter is to see how Bayesian statistics approaches the issue of model selection, to consider the Bayes Information Criterion [BIC] methodology which is placed within Bayesian statistics, to defend the BIC methodology against various objections, and finally to compare and contrast the AIC and BIC methodologies.

4.1 Bayesian Statistics

4.1.1 Bayes Theorem

Bayesian statistics is a unified methodology of statistical inference that is based on Bayes Theorem [cf. section 1.3.1]. Recall the Theorem:

$P(A|B) = P(B|A)xP(A)/P(B)$ where $P(B) > 0$

Let us replace proposition A with proposition H, which stands for 'the hypothesis is true', and proposition B with proposition E – 'a certain amount of evidence has been observed'. Hence:

$P(H|E) = P(E|H)xP(H)/P(E)$ where $P(E) > 0$

P(H|E) is called the posterior probability of H in the light of evidence E [in other words, the probability that H is true after data has been observed], P(E|H) is the likelihood of observing the evidence E conditional on the truth of the hypothesis H [often simply referred to as the likelihood], P(H) is the prior probability of H being true [or the probability of H being true before data has been observed] and P(E) is the probability of observing data mentioned in proposition E. The prior probabilities in Bayesian statistics are always conditional on the background knowledge. So properly speaking we should write P(H|background knowledge) instead of simply P(H). However, we omit the background knowledge to simplify the notation.

In Bayesian statistics, the Bayes theorem is usually expressed as

$$P(H|E) \propto P(E|H)xP(H)$$

That is, the posterior is proportional to likelihood times the prior. The constant of proportionality is 1/P(E).

The process by which we draw inference in Bayesian statistics is the following. We first have a prior probability P(H) of the hypothesis being true in the first place. Then we observe data and work out what the likelihood of it is. Then we update our probability of the hypothesis H in the light of data E through the Bayes Theorem. The prior does not have to be purely a priori. It is in fact conditional on all the available information before we observe the new data sited in proposition E [or the data that we are not aware of as yet]. The important point to note about Bayesian statistics is that once the posterior distribution [or posterior probability density in case of continuous variables] is generated by means of the Bayes Theorem, further inference in it [such as the determination of the highest density region – Bayesian equivalent of the confidence intervals] is solely based on this posterior, that is, on the probability distribution in the light of the current observations. Let us now consider in detail the formation of prior probabilities. We shall not impart similar attention to likelihoods since their calculation is uncontroversial.

4.1.2 Priors

4.1.2.1 Objectivity and the Principle of Indifference

Let us begin by considering the origins of a prior probability distribution. As we already mentioned, the prior distribution [as, indeed, any other probability distribution in Bayesian statistics] reflects the subjective degree of belief of a given researcher about, for instance, distribution of probabilities associated with different values of a parameter. The use of prior distributions is considered by many to be the major weakness of the Bayesian approach [indeed, its Achilles heel]. The charge is that since it is possible for different researchers to come up with widely divergent priors, their posteriors would be quite different as well, thus making the science of

statistics a thoroughly subjective enterprise. This is an unpalatable conclusion if objectivity is something that science should strive for. There have been several proposals over the years [indeed over a couple of centuries] as the possible ways in which the priors can be made more 'objective'. By far the most popular idea has been the Principle of Indifference [POI].[54] The POI states that every basic event in the outcome space should be assigned equal probability. To illustrate, in our die-throwing example we have two basic events - odd and even number on the face of the die. So, by the POI, prior probability of odd number and prior probability of even number should be 0.5 respectively. Unfortunately the POI runs into trouble. If we transform the continuous parameter space in a non-linear way [say, if we have parameter v, we transform it into something like $1/v$], then what was a uniform distribution over v [uniform distribution is the result of application of the POI in case of continuous parameters] becomes a non-uniform one. Here is a nice example:

> 'Suppose we have a mixture of wine and water and we know that at most there is 3 times as much of one as of the other, but nothing more about the mixture. We have $1/3 \leq$ wine/water ≤ 3 and by the Principle of Indifference, the ratio of wine and water has a uniform probability density in the interval [1/3, 3]. Therefore $P(\text{wine/water} \leq 2) = (2–1/3)/(3–1/3) = 5/8$. But also $1/3 \leq$ water/wine ≤ 3 and by the Principle of Indifference, the ratio of water to wine has a uniform probability density in the interval [1/3, 3]. Therefore $P(\text{water/wine} \geq 1/2) = (3–1/2)/(3–1/3) = 15/16$. But the events 'wine/water $\leq$ 2' and 'and water/wine $\geq$ ½' are the same, and the Principle of Indifference has given them different probabilities.'
> Gillies (2000):38

There are other approaches to the 'objective' priors, such as the use of entropy priors[55]. However, in the limit this prior is uniform, and hence does suffer from the POI paradoxes as well as the original POI itself [cf. Howson and Urbach (2006):section 9.a.3].

Subjective Bayesians respond with two arguments. Firstly, said subjectivity of priors is not a weakness of the method, but its strength. Secondly, there are various technical results that can crudely be called 'washing out theorems' that show that under quite general conditions [the most important of which is that the prior assigns

[54] Here we use terminology introduced by Keynes (1921).
[55] cf. Williamson (2007) and (2010)

a non-infinitesimal probability in the region of likelihood], as the number of observations accumulates, the likelihood rapidly gains disproportionately larger weight than the prior, and the posteriors obtained with different priors in the limit converge onto the same value. [56] We shall say more about priors in the next section.

4.1.2.2 Conjugate Priors

Suppose that we have managed to bring ourselves to be happy with the idea that there is no such thing as objective priors [for a lot of people this happiness is unreachable]. How are we to build our prior distribution then? Let us use the example of throwing the die and noting the even and odd numbers. Theoretically our prior can be of any shape [naturally subject the constraints given by the probability axioms]. However, if the prior comes from a different family of distributions to that of the posterior, our calculations would be rather difficult. So quite often in practical applications so-called conjugate priors are used. That is, conjugate priors are such that they come for the same family of distributions as the posterior. Naturally, one would not want to sacrifice the ability of express one's beliefs for sheer mathematical convenience. Very often, however, conjugate priors are flexible enough to allow one to express one's prior degrees of belief sufficiently well.

So, back to rolling the die. As was the case in the section of classical statistics, suppose that we are happy that we have the Binomial set up. A conjugate prior for a Binomial is a Beta distribution.

A random variable X has a Beta distribution if its p.d.f. is:
$x^{a-1}(1-x)^{b-1}/B(a,b)$, $0 < x < 1$,
where $B(a,b)$: $B(a,b)=\int x^{a-1}(1-x)^{b-1}dx$ (integrated from 0 to 1)
The mean and variance of X: $E(X) = a/(a+b)$, $Var(X) = ab/(a+b)^2(a+b+1)$

Below are some diagrams showing some examples of Beta distribution plots, where 'First' stands for the parameter *a*, and 'Second' - for *b*[57].

[56] Howson and Urbach (2006):chapter 9
[57] The diagrams have been generated with the MiniTab software.

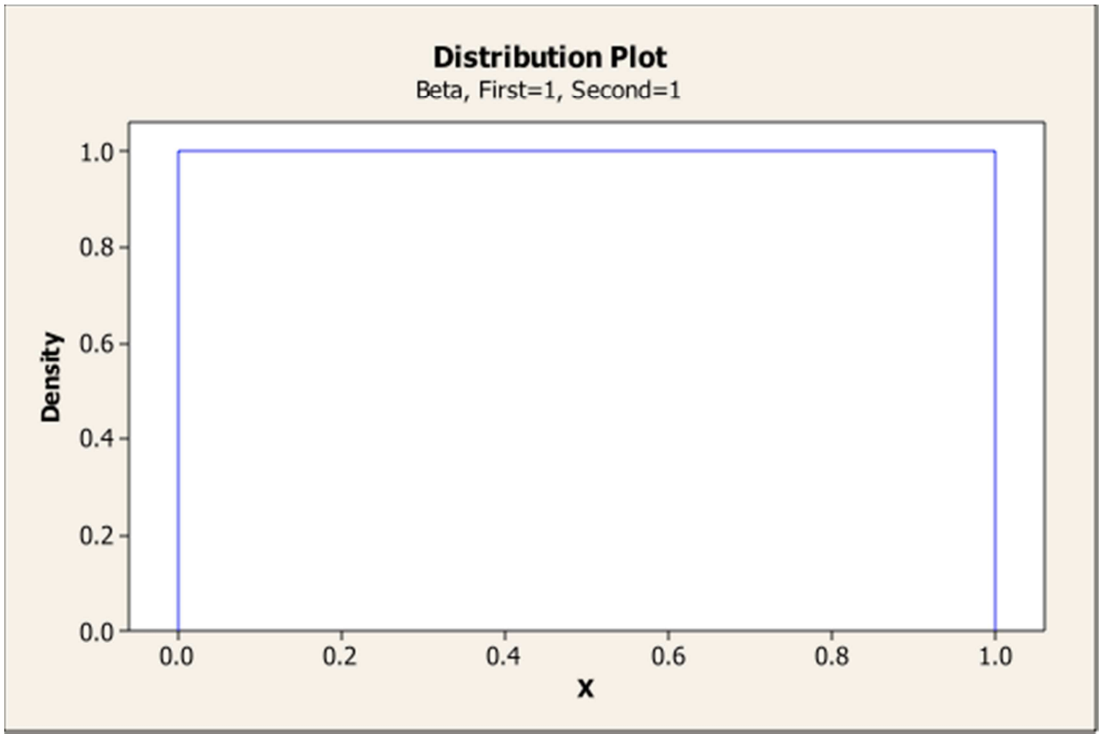

Diagram 2

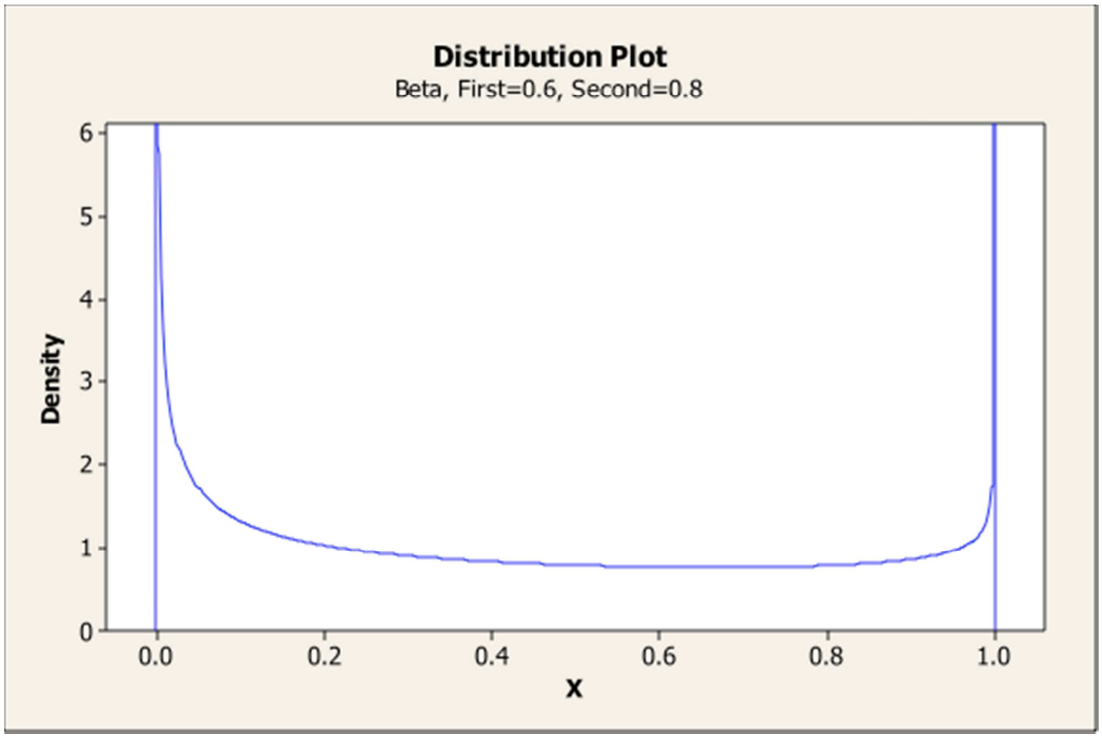

Diagram 3

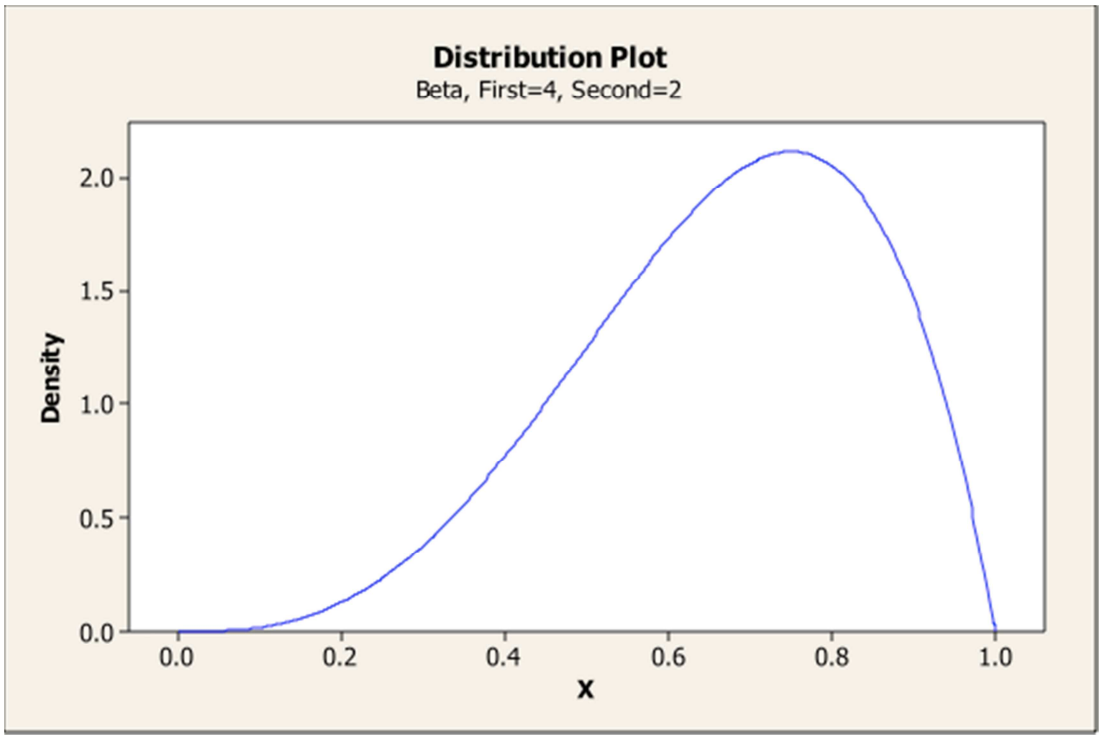

Diagram 4

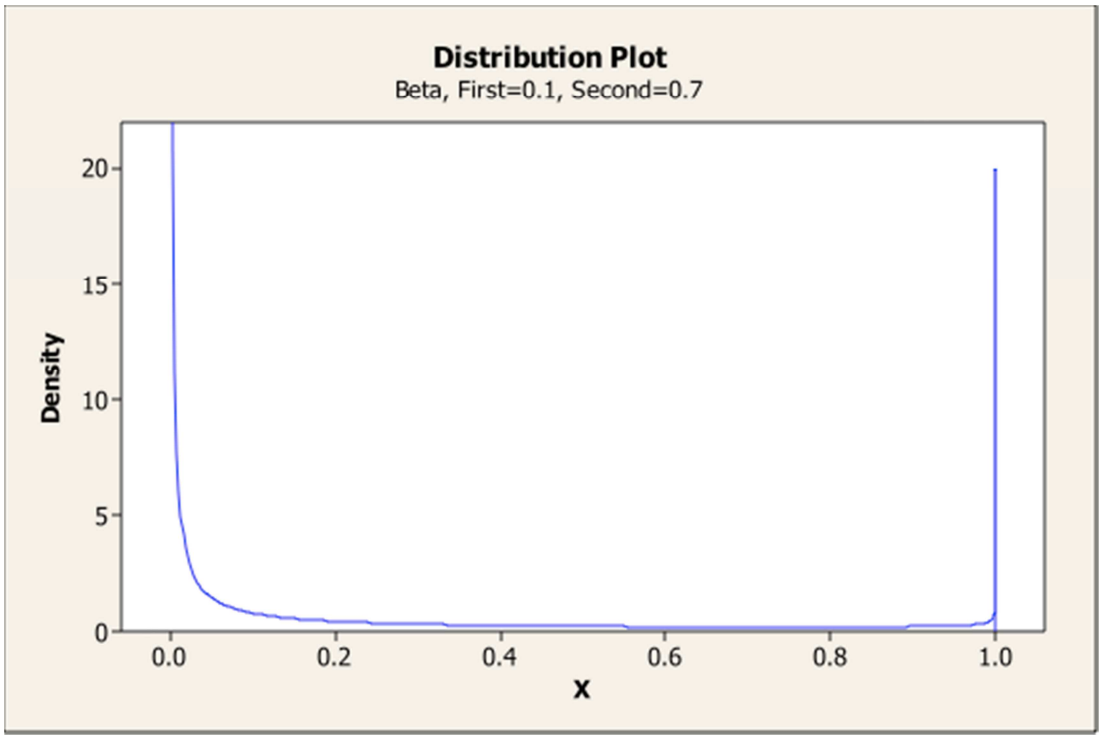

Diagram 5

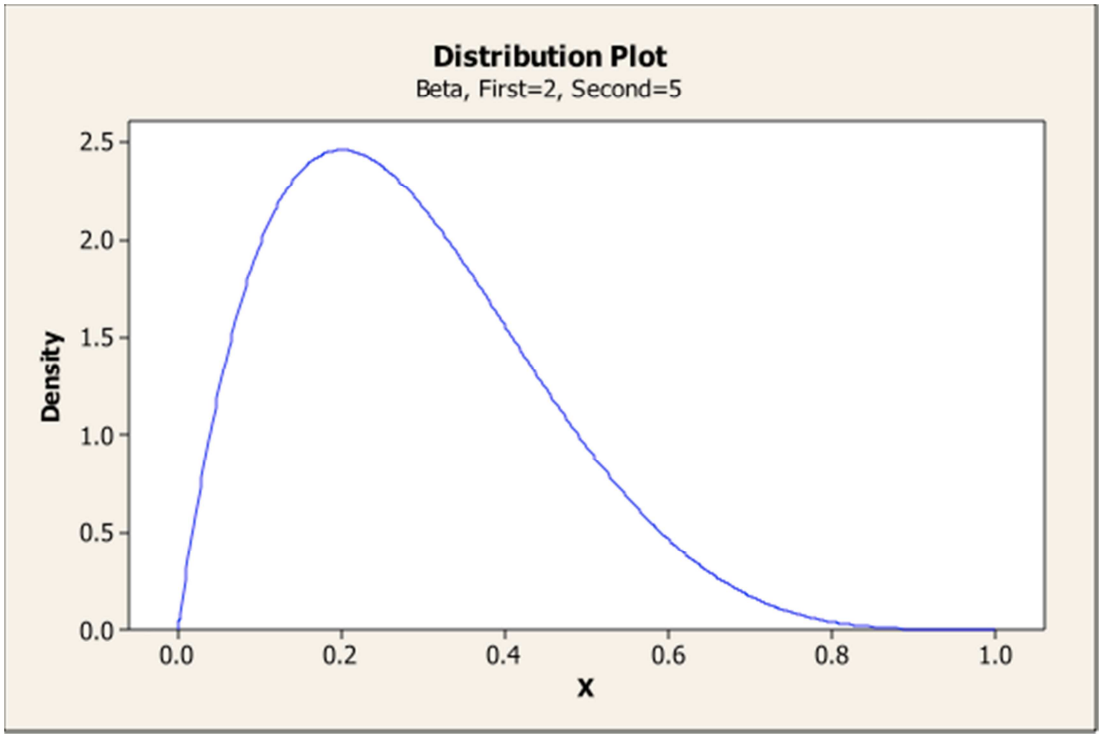

Diagram 6

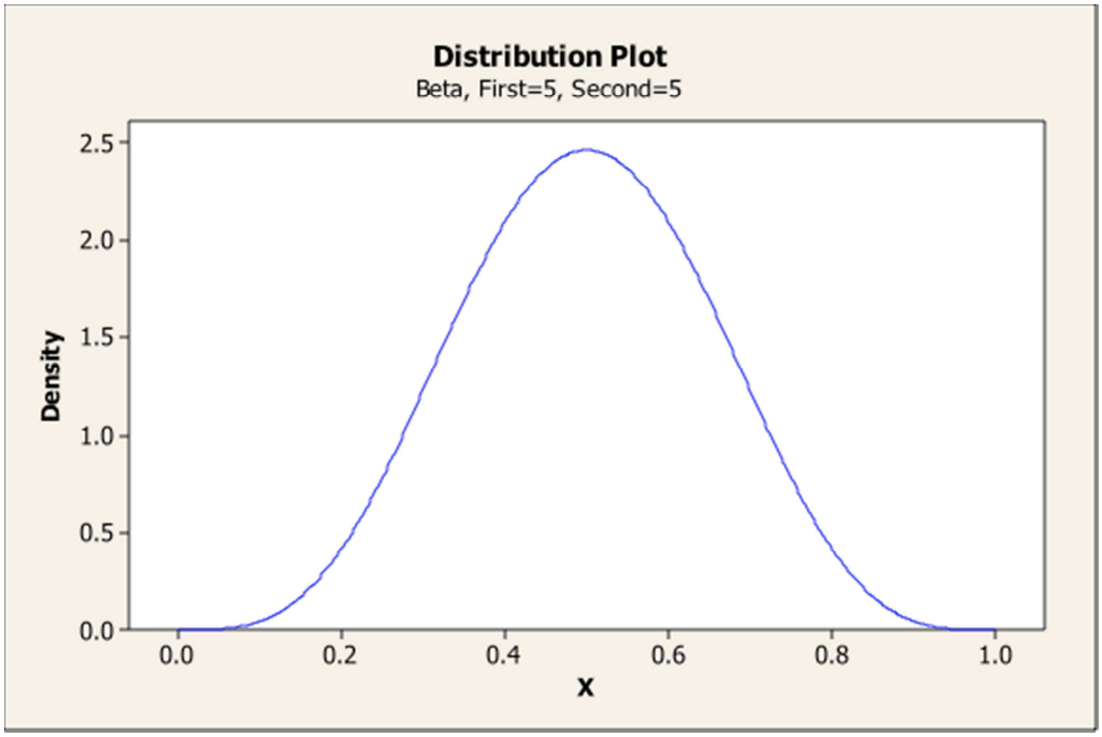

Diagram 7

The diagrams allow visual appreciation of a high degree of flexibility with which one's prior probability distribution can be expressed using the Beta distribution. In particular, when $a = b$, the corresponding p.d.f. is symmetric [diagram 2 is a special case of the uniform distribution [when $a = b = 1$] and diagram 7 looks similar to the Normal distribution with $a = b = 5$; generally when $a = b$ = constant, as the constant grows larger, the distribution concentrates around the middle values with increasingly smaller variance]. When both a and b are smaller than 1 then the distribution is almost bi-valued, that is, most of the probability density is distributed in the extremes of the distribution, rather than in the centre [see diagrams 3 and 5]. In particular, when $a < b$, there is more density on the left hand side, and the greater the difference between a and b, the more density there is on the left hand side. The opposite holds when $b > a$. Finally, on diagrams 4 and 6, both a and b are greater than 1. On the diagram 4 $a > b$, hence the distribution is skewed to the right, whereas when $b > a$ the opposite holds [see diagram 6]. The severity of the skew depends on the magnitude of the difference between a and b.

The prior distribution is $p^{a-1}(1-p)^{b-1}$
The likelihood is $p^{r}(1-p)^{n-r}$, where p is the probability of success, r is the number of successes out of n observations.
So, the posterior probability is proportional to $p^{a+r-1}(1-p)^{b+n-r-1}$
The posterior distribution is $p^{a+r-1}(1-p)^{b+n-r-1}/\mathrm{B}(a+r,b+n-r)$, $0<p<1$
So, for instance, if we have a uniform prior, i.e., $a = b = 1$, the posterior distribution is $p^{r}(1-p)^{n-r}/\mathrm{B}(r+1,n-r+1)$

4.1.3 Model Selection Based on Bayes Factors

Suppose that we have two point hypotheses H_0 and H_1 [just as in chapter 2] that we would like to compare in the light of observed data. In the Bayesian approach it is done on the basis of the Bayes factor. The easiest way to define the Bayes factor is in terms of the odds ratio. Posterior odds is the ratio of posterior p_0 probability of H_0 to the posterior probability p_1 of H_1 – that is, p_0/p_1. Prior odds is the ratio of prior probability c_0 of H_0 to the prior probability c_1 of H_1 – c_0/c_1. So, Bayes factor (B) in favour of H_0 against H_1 is the ratio of the Posterior odds:

$B = (p_0/p_1)/(c_0/c_1)$

Good (1950) proposed an interpretation of B such that $0 < B < 1$ means that H_1 is favoured in comparison to H_0, $1 < B < 10$ means that H_0 is moderately favoured to H_1, $10 < B < 100$ – that H_0 is strongly favoured to H_1.

Let us apply this reasoning to an example that we used in the subsection on Neyman-Pearson methodology. There we had a die-rolling set up such that we set out to roll the die 120 times in order to test two hypotheses against one another, viz., that the probability of obtaining even outcomes is 0.55 [denote it as H_0] or 2/3 [denote it as H_1]. Suppose, as we did, that we observe 70 even numbers out of 120 rolls of the die. What inference would we draw within the Bayesian methodology?

In the previous subsection it was noted that the likelihood in this set up is $p^r(1-p)^{n-r}$, so here it is $p^{70}(1-p)^{50}$. Let us use the conjugate prior in the form of Beta distribution. Hence our prior is $p^{a-1}(1-p)^{b-1}$, where we should determine the value of parameters a and b. Suppose that we opt for a uniform prior $a = b = 1$. The posterior probability is proportional to $p^{a+r-1}(1-p)^{b+n-r-1}$, so in our case it is $p^{70}(1-p)^{50}$. Let us calculate the Bayes factor B: $B = 0.55^{70}\text{x}0.45^{50}/0.66^{70}\text{x}0.34^{50} = 4.6608$ (4 d. p.) Now suppose that we had a different prior, say where $a = 3$ and $b = 2$. Let us calculate the Bayes factor for this eventuality: $B = 0.55^{72}\text{x}0.45^{51}/0.66^{72}\text{x}0.34^{51} = 4.2826$ (4 d. p.) The Bayes factors are very similar. On the basis of I. J. Good's interpretation of Bayes factors, this implies that H_0 is moderately favoured over H_1. Note that on the basis of this quite moderately sized set of data, the Bayes factor is not that sensitive with respect to the priors – it is dominated by the likelihoods.

4.1.4 Point Estimation and Bayesian Confidence Interval

Very often Bayesian point estimates are biased in the Classical sense [cf. section 2.4.1] and are different to the Classical estimators. For example, in the Binomial case the Bayesian point estimate for the probability of success is: $(a+r)/(a+b+r)$. If the prior is uniform, this corresponds to $(r+1)/(n+2)$. In fact, the only prior that corresponds to the Classical estimate of r/n is when $a = b = 0$.

Bayesian confidence intervals are often [but not always] the same as those derived in Classical statistics, but their interpretation is quite different. That is, unlike their classical counterparts [cf. section 2.5] the Bayesian confidence intervals are interpreted directly in terms of probabilities. So, to say that a parameter lies within a certain interval with 95 % confidence is to say that the parameter has a 95 % probability of lying within said interval.

4.2 Bayes Information Criterion

In fact the name 'Bayes Information Criterion' is somewhat misleading since what has come to be widely known as BIC neither has anything to do with Shannon's Information Theory [cf. Shannon and Weaver (1949)], nor is it *the* one and only Bayesian Information Criterion. Indeed, there is a plethora of model selection criteria within Bayesian framework – cf. Spiegelhalter (2002), Konishi and Kitagawa [(2008):chapter 9]. However, we will concentrate on BIC in particular, for three reasons. Firstly, even though BIC and other Bayesian criteria differ in details, they remain based within the same methodology, so the philosophical points by and large apply to them all. Secondly, BIC has received particular attention in the philosophical literature. Thirdly, like AIC, BIC is the most widely used Bayesian model selection criterion in statistical practice.

BIC is due to Schwarz (1978). That is why it is sometimes referred to the SIC [Schwarz Information Criterion] or the Schwarz Criterion or Schwarz's Bayesian Information Criterion. However, we shall continue referring to it as BIC following the most common usage in the literature.

To get started, recall the discussion in section 4.1. One of the features of Bayesian statistics that we looked at there were the Bayes factors. Bayes factors are used to see which model from a given range is favoured by the sample data at hand. Bayes factors are the basis of model selection in Bayesian statistics.

Recall that if we suppose that we have two models, say, H_1 and H_2, and for simplicity of exposition assuming that H_1 and H_2 are mutually exclusive and exhaustive, so that our data E arose from one of these models, the following holds:

$$\frac{pr(H_1 \mid E)}{pr(H_2 \mid E)} = \frac{pr(E \mid H_1)}{pr(E \mid H_2)} \times \frac{pr(H_1)}{pr(H_2)}$$

That is, Posterior odds = Bayes factor x Prior odds. If the priors on our models are equal, as it is assumed in BIC, then the Posterior odds = Bayes factor. Supposing that we are using continuous random variables, then *pr* is a probability density. Then:

$$B_{12} = \frac{pr(E \mid H_1)}{pr(E \mid H_2)} = \frac{\int pr(E \mid \theta_1, H_1)\pi(\theta_1 \mid H_1)d\theta_1}{\int pr(E \mid \theta_2, H_2)\pi(\theta_2 \mid H_2)d\theta_2}$$

where θ_i is a *K*-dimensional vector of parameters of the model H_i, and $\pi(\theta_i|H_i)$ is the prior probability of the vector of parameters θ_i given the model H_i – so called parameter prior. In order to obtain the full Bayesian solution the Bayes factor has to be combined with the model priors. Thus there are two prior distributions involved.

The integrals involved in the above expression for B_{12} are often mathematically intractable, and have to be estimated by numerical methods such as Laplace approximation [cf. Kass and Raftery (1995): 777-778, and Konishi and Kitagawa (2008): 231-236]. Bayes Information Criterion is essentially an easy-to-calculate approximation to the natural logarithm of the Bayes factor.

In a large sample with independent identically distributed data points the following holds [for a full formal proof cf. Schwarz (1978), Cavanaugh and Neath (1999) or Burnham and Anderson (2002)]:

$$-2\ln pr(E \mid H_i) = -2\ln\{\int pr(E \mid \theta_i, H_i)\pi(\theta_i \mid H_i)d\theta_i\} \approx BIC_i$$

where $BIC_i = -2\ln(E \mid \hat{\theta}_i, H_i) + K_i \ln n$, and $\hat{\theta}_i$ is the maximum likelihood estimator of the K_i–dimensional parameter vector θ_i of the model H_i. In the notation of chapter 2, BIC = $\mathbf{-2\ln(L(\hat{\theta} \mid data, g_i)) + Kln(n)}$

Notice that $-2\ln B_{12} \approx BIC_1 - BIC_2$

The BIC methodological rule is the same as for the AIC – that is, to choose the model which has the smallest BIC score.

The following table shows how the differences in BIC scores between two mutually exclusive exhaustive models correspond to the differences in Bayes factors and posterior probabilities [on the assumption of equal model priors]. This table resembles a similar 'rule of thumb' used in the AIC methodology [cf. section 3.2.2].

Grades of evidence corresponding to Values of the Bayes Factor for H_1 against H_2, the BIC Difference and the Posterior Probability of H_1

BIC Difference	Bayes Factor	*pr* (H_1|E)	Evidence
0 – 2	1 – 3	0.5 – 0.75	Weak
2 – 6	3 – 20	0.75 – 0.95	Positive
6 – 10	20 – 150	0.95 – 0.99	Strong
> 10	> 150	> 0.99	Very Strong

From Raftery (1995):138, notation modified to fit our usage

We shall provide an example of use of the BIC in section 4.5.2. In fact it shall be the same Hald's Cement Hardening Data example that we employed in section 3.2.2.

4.3 Philosophical Issues with BIC

4.3.1 Nesting

Let us return to models LIN and PAR that we used as examples to introduce the AIC in chapter 3. LIN is a model which has all linear equations as its elements: $y = a + bx$. PAR is a model which has all parabolic equations as its elements: $y = a + bx + cx^2$. As things stand, LIN is a proper subset of PAR, i.e., every element of LIN is an element of PAR, and PAR has more elements than LIN. LIN is said to be *nested* in PAR. Hence according to probability calculus $pr(\text{LIN}) \leq pr(\text{PAR})$ [Popper (1968)]. Probability of LIN equals to probability of PAR just in case all the elements of PAR in which $c \neq 0$ have probability zero. The same inequality applies to the posterior probabilities of LIN and PAR. That is, for any data E, the following holds: $pr(\text{LIN}|E) \leq pr(\text{PAR}|E)$. So the posterior probability of PAR is at least as large as that of LIN whatever evidence we observe. Hence the posterior odds ratio $[pr(\text{LIN}|E)/pr(\text{PAR}|E)] \leq 1$, so in Bayesian methodology LIN would not be preferred to PAR on any evidence at all. However, on the basis of the BIC methodology it is possible to prefer LIN to PAR. That is, it can be the case that [BIC(LIN) – BIC(PAR)] < 0. This leads Forster (2000):214 to a conclusion that 'Bayes' method is one thing and BIC is another. The latter is not always an approximation of the former.' Let us see where the difference between Bayes' method and BIC lies in this case.

Recall from section 4.2 of this chapter that the BIC method provides an approximation to the Bayes Factor B_{12}. The Bayes factor is essentially a ratio of integrated likelihoods of the data, which is an average of the likelihoods assigned to the data by each element of the model weighted by the prior probability distribution over all the elements of the model given that the model is correct. Even though LIN is nested in PAR, their Bayes factor is not restricted to any particular interval of values. That is, the Bayes factor of LIN against PAR ($B_{LIN,PAR}$) can be greater than 1. Intuitively this is because the prior probability distribution over the parameters in PAR is spread more 'thinly' over the three parameters rather than over the two, as it is the case in LIN. If the data exhibits considerable linearity then the likelihoods of the elements in LIN are weighted higher by their priors within the integrated

likelihood than their linear counterparts in PAR. [cf. Kuha (2004):213] Hence the possibility of $B_{LIN,PAR} > 1$. In general, for any model H_1 that is nested in another model H_2 the following holds. If H_1 contains an element which is 'true' (we shall spend more time on the topic as to what is a 'true' model in the next subsection) and thus H_2 contains the true element as well, then as the number of observations tends to infinity, B_{12} also tends to infinity. This is the case for almost any distribution of prior probabilities to the parameters given the respective models [cf. Dawid and Senn (2011):19].

Recall, however, that fully Bayesian model selection is based on posterior odds, where Posterior odds = Bayes factor x Prior odds. If we base our model selection solely on Bayes factors, our model selection is not affected by the issue of nesting of models, but our methodology is semi Bayesian, because we only use the priors over the parameters given the correctness of respective models, but do not employ priors over models themselves. Once we combine a Bayes factor with prior odds we obtain the result that the posterior odds of LIN against PAR are never greater than one. The BIC method provides an approximation of Bayes factor, so it also provides an approximation to the posterior odds just in case the priors over models themselves are equal, i.e., the prior distribution of the models is uniform – the Prior odds are then equal to 1. So the fully Bayesian model selection based on nested models would never favour a model with fewer adjustable parameters to an alternative with more adjustable parameters. As we discussed in the beginning of chapter 3, this is not a desirable feature of a model selection methodology.

Forster's conclusion cited above is correct in the case of nested models – if we wish to do fully Bayesian model selection properly we cannot work with nested models. A natural solution to this issue seems to present itself. Once we remove all linear elements from PAR and thus define PAR*: $y = a + bx + cx^2$ where $c \neq 0$, then LIN is no longer nested in PAR*, and there are no longer any restrictions on what values both prior and posterior odds can take. Moreover, surely it is more fruitful to select among incompatible models rather than between general models and their special cases [cf. Howson and Urbach (2006):289].

Nonetheless, there are further arguments that the move from, in Forster's terminology, truly nested models (like LIN and PAR) to quasi-nested models (like LIN and PAR*) makes the Bayesian model selection somehow inferior to the other methods which do not have this issue: 'This maneuver succeeds in restoring consistency to [Bayesian] claims. Nevertheless, it does not resolve the puzzle about why there *should* be any difference between truly nested and quasi-nested models. In the other methods of model selection, such as AIC …, there is no difference between these two cases.' [Forster(2000):214]

Curiously we have not come across the following considerations being made explicit in the extensive literature on model selection. Let us investigate as to why the AIC methodology works equally well with both nested and non-nested models. Let us use models LIN and PAR again. To calculate the AIC scores we find an element of each model which has the maximum likelihood within the respective model. Within LIN that would obviously be a particular line. What about the element which has the maximum likelihood within PAR? It would almost invariably be a parabola with $c \neq 0$ [unless all the data points lie on a straight line, in which case the element with the maximum likelihood will be the same in both LIN and PAR. In the realm of probabilistic statistical modelling that we are concerned with we would expect this eventuality to be extremely rare.]. A parabolic curve has three adjustable parameters rather than two as it is the case for a linear curve, hence allowing the former to fit the data better, and thus to have a higher maximum likelihood. So, even though LIN is nested within PAR, as far as using AIC for model selection is concerned, PAR would almost always be represented by a parabola and penalised for using three adjustable parameters. The fact that LIN is nested in PAR is therefore irrelevant – LIN and PAR* would always yield exactly the same AIC scores as their nested counterparts [bar the case of complete linearity in data]. AIC-based model selection would have exactly the same outcome whether the models in the choice set are nested or not. This result generalises to nested models of any mathematical structure. By using only incompatible models in our choice set we can use both AIC and BIC at the same time and compare their results.

In our view the puzzle as to why the move from truly nested to quasi-nested models in the choice set should make a difference is answered rather simply in the light of

the discussion in this section. It makes a difference in the case when we wish to use BIC methodology in the fully Bayesian way. We think that 'quasi-nested' terminology makes the move from LIN and PAR to LIN and PAR* in the choice set sound insubstantial whereas it is a rather important move. After all, by taking LIN out of PAR, we remove an uncountably infinite subset of PAR, which is not that trivial. Another important move, once the non-nesting of the models in the choice set is established, is the assumption of a uniform prior over the models in order for the differences in BIC scores to directly approximate the posterior odds on models. We shall look further into this assumption in section 4.4.1.

4.3.2 Truth

There are two closely related issues that have been identified with regards to the BIC methodology and truth.

Firstly, it is often argued [for example, Spiegelhalter et al. (2002)] that in order for the BIC to perform properly as a model selection criterion it is necessary to have a "true" model in the choice set. In this context by the "true" model it is usually meant something along these lines: "a model *precisely* representing the full reality underlying the phenomena in question". Within the AIC methodology a true model is such that its Kullback-Leibler divergence from the putative "truth" is zero. It seems rather unlikely that every time that we choose models to constitute the choice set we manage to include a true one in it. So in what are no doubt numerous cases when there are no true models in the choice set the application of the BIC methodology seems meaningless and inappropriate.

Secondly, it is said [for example, Forster and Sober (1994):22] that AIC and BIC were designed for different purposes. Namely, AIC was designed to maximise predictive accuracy and BIC to maximise the probability of a model to be true. So, they are best for the respective jobs they were designed for, and no more.

Indeed, the original derivation of the BIC due to Schwartz (1978) contains an assumption that the true model is within the choice set. However, since then the BIC has been derived in a more general way without employing the true model

assumption – cf. Cavanaugh and Neath (1999). Given this, a question naturally arises as to what we are to make of model probabilities within the BIC methodology. There is a mathematical theorem which states that for independent identically distributed sampling as the number of observations *n* tends to infinity one of the models within the choice set tends to 1 and the rest tend to 0 in probability [Burnham and Anderson (2004):276]. What are we to make of this result? What does $pr\ (H_i|E) = 1$ mean in the case when no model in the choice set is true?

We can say that a model is *quasi*-true if it is the closest model to truth in the Kullback-Leibler sense in the choice set. The asymptotic convergence in probability to 1 of one of the models within the choice set means that this model is quasi-true in the sense indicated [Burnham and Anderson (2004)]. It is curious to see the K-L divergence emerging in the Bayesian context of the BIC methodology. Nevertheless here it is. There is actually another interesting way this connection works via scoring rules.

Scoring rues are designed to measure predictive performance against observations of probabilistic models [both theoretical and statistical as per distinction introduced in chapter 1] or of probability judgements expressed by individuals. Here, as in the rest of thesis, we shall concern ourselves with probabilistic predictions derived from models. As usual, it is perhaps most illuminating to explain the concept by means of an example. [For a rigorous overview of scoring rules cf. Gneiting and Raftery (2007).] Suppose that we have two models H_1 and H_2 which provide probabilistic predictions of whether it will rain on a given day. Suppose that we would like to have a comparison of their predictive performance by means of using a mathematical rule which quantifies a discrepancy between the probabilities that the models yield of it raining next day and the actual observations of the events. In the table below [which is a stylised version of the table in Baron (2008):120] in the top row denoted 'Event occurred?' 'Yes' stands for the observation that it rained the next day after the models provided probabilistic forecasts, and 'No' stands for the event that it did not rain. In each column there are probabilities of it raining on the given day provided by each of the two models respectively.

	Event occurred ?	Yes	No	No	Yes	Yes	Logarithmic Total Score	Quadratic Total Score
Probability of event occurring given by:								
H_1		0.9	0.1	0.4	0.8	0.3	-2.14866	0.71
H_2		0.8	0	0.3	0.9	0.1	-2.98776	0.95

Now that we have data what formal expression should we use to measure the predictive performance? One of the popular scoring rules is the quadratic rule. It works the following way. Let us take the first 'Yes' column in the table above as an example. There model H_1 predicted rain with probability 0.9 and model H_2 with probability 0.8. Since it did actually rain we take the 'true' probability of it raining on that day to have been 1. [As we mentioned in chapter 1 when introducing the elementary probability theory, it is not an aim of this thesis to delve into the issue of interpreting probabilities.] In the quadratic rule we square the discrepancy between the 'true' probability and the predicted probability. So, for the day in the first column the discrepancy for model H_1 is $(1 – 0.9)^2 = 0.01$; for model 2: $(1 – 0.8)^2 = 0.04$. The total quadratic score is provided by adding all of the discrepancies together thus:

Quadratic Total Score for model $H_1 = (1 – 0.9)^2 + (0 – 0.1)^2 + (0 – 0.4)^2 + (1 – 0.8)^2 + (1 – 0.3)^2 = 0.71$

Quadratic Total Score for model $H_2 = (1 – 0.8)^2 + (0 – 0)^2 + (0 – 0.3)^2 + (1 – 0.9)^2 + (1 – 0.1)^2 = 0.95$

The model with the lowest quadratic total score is considered to be the most predictively successful for a given sample of data. The minimum achievable total quadratic score is zero. In fact within the theory of scoring rules the quadratic rule is identified as a *strictly proper* rule. Informally [for the formal definition cf. Gneiting

and Raftery (2007):359], strictly proper rules are such that there is no strategy of assigning probabilities to events in order to improve the total score (in the quadratic rule's case that would mean to lower the total score) except than to stick to the probabilities that a given model issues. That is, there is no way to 'beat the system', in a manner of speaking. Strictly proper scoring rules bear a certain similarity to the exclusion of gambling systems in the context of gambling. A gambling system is a set of instructions specifying when and how much to bet when playing a game of chance [for example, roulette] with the aim of improving monetary gain – 'beating the odds' [for an in depth consideration of the law of excluded gambling systems cf. Gillies (2000):chapter 5]. Baron (2008):121 gives an example of an improper scoring rule.

Another example of a strictly proper rule, which is in fact pertinent to our topic of the BIC methodology is the logarithmic scoring rule. It works in the following way. If a model predicts the occurrence of an event with probability p and the event subsequently occurs, then the score is $\ln(p)$. If the event does not occur, then the score is $\ln(1 - p)$. So, for the day in the first column the logarithmic score for the model H_1 is ln(0.9) = -0.10536; for H_2: ln(0.8) = -0.22314. The total logarithmic score is provided by the sum of the individual scores. Hence:

Logarithmic Total Score for Model H_1 = ln(0.9) + ln(1-0.1) + ln(1-0.4) + ln(0.8) + ln(0.3) = ln(0.9x0.9x0.6x0.8x0.3) = -2.14866

Logarithmic Total Score for Model H_2 = ln(0.8) + ln(1-0) + ln(1-0.3) + ln(0.9) + ln(0.1) = ln(0.8x1x0.7x0.9x0.1) = -2.98776

The model with the highest logarithmic total score is considered to be the most predictively successful for a given sample of data. The maximum achievable total logarithmic score is zero.

Good (1952) points to the following result:
$\ln(BF_{12})$ = total logarithmic score of model 1 – total logarithmic score of model 2, which with simple algebraic manipulations is approximated by $-0.5x(BIC_1 – BIC_2)$.

So there is a way to interpret BIC scores as providing a measure of predictive success, on par with the AIC methodology. Another striking result is that the mathematical expectation of a logarithmic score is equal to the Kullback-Leibler divergence [Ehm and Gneiting (2009, Addendum 2010):4].

Recall that the AIC methodology aims to provide an *unbiased* estimate of the expected relative K-L divergence to 'truth'. Then it seems surprising that even though the BIC methodology also has a link to the K-L divergence, the numerical expressions of the AIC and BIC criteria are different. In a nutshell, the difference lies firstly in the use of maximum likelihoods in the AIC as opposed to the integrated likelihoods in BIC and secondly in the fact that the penalty term $2K$ appears as a correction of a bias [in the sense that this notion is explained in chapter 1] in the AIC whereas $K\ln(n)$ in the BIC appears during approximation of the integrated likelihood. The connection between AIC and BIC is explored further in section 4.4.

4.4 Connection between BIC and AIC

Perhaps it does not come as a huge surprise that AIC and BIC are connected. After all, the only difference in the formal expressions between AIC and BIC is that the penalty term K [i.e., the number of adjustable parameters] is multiplied by 2 in AIC and by $\ln(n)$ in BIC. There are two ways in which we shall explore this connection. Both of these shall show what would be required in order to yield the AIC from the Bayesian perspective of BIC. This is the easiest way to exhibit the link between AIC and BIC, since Bayesian methodology allows us the flexibility of priors. Recall that in order to use BIC as an approximation to the fully Bayesian way, two sets of priors are determined – the priors over parameters given the models, and the priors over models themselves. In two subsections below we shall explore the kind of priors required to yield AIC from BIC. In section 4.4.1 we shall look at the type of model prior required [while using the same parameter prior as in BIC] in order to derive AIC. In section 4.4.2 we shall look at the type of parameters prior required [while using the same uniform prior over models as in BIC] in order to derive AIC.

4.4.1 Connection via Model Priors

Burnham and Anderson (2004) show that if we use the following model prior instead of a uniform one, we derive the AIC rather than BIC:

$$q_i = \frac{\exp(\frac{1}{2}K_i \ln(n) - K_i)}{\sum_{r=1}^{R} exp(\frac{1}{2}K_r \ln(n) - K_r)}$$

This prior is an increasing function of both of the size of data sample and of the number of adjustable parameters. That is, for a given number of observations in the sample, models with relatively larger number of adjustable parameters have higher probabilities than models that have relatively fewer number of adjustable parameters. Also an increase in the sample size brings about an increase in the difference in probability of models with different numbers of adjustable parameters. This can be seen in the simple example in the table below, where we performed calculations of such prior probabilities of two models with two and three adjustable parameters respectively with samples consisting of 10 and 100 observations respectively. The two models are assumed to be exhaustive and mutually exclusive.

	$n = 10$	$n = 100$
$K = 2$	0.4626	0.2137
$K = 3$	0.5374	0.7863

Burnham and Anderson call the model prior which takes us from the BIC to the AIC a ‘savvy’ prior and argue that this prior is more sensible than the uniform prior used in BIC. In fact they go as far as to state that the very use of the uniform model prior implies that the model selection is done in order to find the true model rather than in order to maximise the predictive performance. Unfortunately they do not offer any argument as to why this should be the case. We disagree with their position. In our view any model prior whatsoever expresses the probability assignment to each model in the choice set that it is (quasi-)true given the background knowledge in the domain of inquiry. The model prior does not and cannot by itself express our belief [or lack thereof] that the choice set contains a true model. For any model prior whatsoever we can represent the posterior odds as the difference in logarithmic

predictive scores. Moreover, there is another way to show that Burnham and Anderson's claim with regards to the "meaning" of model priors is incorrect. We shall consider it in the next section.

There is a way, however, to argue for Burnham and Anderson's contention that the savvy model prior is more sensible than the uniform one. In fact Popper (1968, Appendix viii) provides a version of such an argument. Popper argues that simpler hypotheses ['simpler' in the precise sense that they have relatively fewer adjustable parameters] have relatively lower probabilities. In his view simpler hypotheses have more empirical content, which is measured by the degree of their testability. Simpler hypotheses are more testable in the sense that there is a greater variety of observations that would falsify them. That is in Popper's view there is a larger number of possible data points that would be incompatible with a simpler hypothesis, and so more possibilities for the simpler hypothesis to be wrong relative to a more complex hypothesis.

> 'Simple statements, if knowledge is our object are to be prized more highly than less simple ones *because they tell us more; because their empirical content is greater; and because their better testable*'.
> Popper (1968):142, original italics

Jeffreys (1961) holds the opposite view to Popper on the issue of probability of relatively simpler hypotheses. In his opinion the simpler the hypothesis is, the higher its prior probability, ceteris paribus. This he calls the Simplicity Postulate. He gives two reasons for this postulate. Firstly, simpler hypotheses are more likely to be predictively successful [Jeffreys (1961):4]. Secondly, the Simplicity Postulate fits well the common scientific practice, at least in physics. That is, Jeffreys argues that physicists behave as if they consider simpler hypotheses more likely to be true by always considering a linear hypothesis first, and only then a quadratic one, and so on [Jeffreys (1961):47 and Jeffreys (1973):63].

Starting with the second of Jeffreys' reasons, our view is that it is perfectly compatible with physicist's behaviour to think that she considers simpler hypotheses first for ease of calculations and in an exploratory way, rather than necessarily due to believing that the simpler hypotheses are true. The order in which a scientist

considers hypotheses does not necessarily imply any particular order of probabilities. Indeed, as we just have seen, Popper reached the opposite conclusion, and his approach fits this scientific behaviour as well as that of Jeffreys.

The first reason for adopting the Simplicity Postulate [that simplicity is the guide to predictive success] requires an independent argument for it. As it stands, it is just an assertion. Prima facie, it would be equally reasonable to state that complexity is the guide to predictive success. It is true that simplicity has for a long time been considered to be one of the attributes of a good scientific theory [cf. for instance Kuhn (1977)]. However, we do not think that adopting the Simplicity Postulate as the constraint on setting the model priors is a sensible strategy. Scientists should be free to set the model priors in the way that they deem appropriate given the particular background knowledge and the domain of inquiry. Note that even though Jeffreys' Simplicity Postulate implies that the prior probability over the models in the choice set is a decreasing function of the model complexity as measured by the number of adjustable parameters that the model contains, still in his own examples he uses the uniform prior over models – "for calculation". We shall consider Popper's argument that the simplicity of a hypothesis varies in the opposite direction to its probability in detail in chapter 5.

Notice, incidentally, that we have so far managed to avoid talk of simplicity, parsimony and such-like notions. In our view, the interpretation of the penalty terms in both AIC and BIC as 'simplicity in action' is unnecessary. In both AIC and BIC the penalty for complexity arises from the formalism itself – in the AIC the penalty term for the number of parameters arises as the correction term for the asymptotic bias, and in the BIC it arises during the process of approximation to the integrated likelihood. The notion of simplicity was not input into either of these methods – it emerged from the formalism as a by-product. Thus we do not concentrate our attention on this feature, for we get no epistemic purchase on it over and above the model selection criteria themselves.

For us there is no full proof formal way to prescribe how model priors should be set. Each particular case demands deliberation on this issue. Every purely formal rule for setting priors is *ad hoc*.

4.4.2 Connection via Parameter Priors

We mentioned in the beginning of this section that there are two ways of deriving the AIC result from the BIC methodology. The first way was to keep the parameter priors the same as in the BIC and to derive a model prior which would take us to the AIC result. This is what we did in the previous section. Now we shall keep the uniform model prior fixed, and show that there is a parameter prior which again takes us to the AIC result from the BIC setting.

This section closely follows Kieseppä (2001a). His approach is to consider how informative any given probability distribution is. From chapter 1 recall that it is often possible to fully determine a probability distribution by two numbers [depending on the distribution] – by its mean and its variance [this is the case for the normal distribution – cf. section 1.3.2]. The variance is the measure of dispersion of a given distribution. That is it measures how spread out the possible values that the parameter can take given the structure of the distribution. The higher the variance the more spread out the distribution is around its mean value. So the variance is said to measure the informativeness of a given distribution in the sense that the higher the variance the less informative the distribution is since there are more possible values that the parameter can take. In the multiple regression case the variance is substituted by the covariance matrix, but the idea is the same. It is also noted that the informativeness of a probability distribution is proportionate to the number of observations. That is the more observations it is based on, the higher its informativeness. It is then possible to rank different probability distributions by their informativeness in terms of the number of observations expected to be required in order to obtain given variance. [For formal treatment of this topic cf. Kieseppä (2001a).] Here is the formula for a general Bayesian model selection criterion without assuming any particular parameter prior:

$$\mathbf{-2\ln(}L(\widehat{\theta} \mid data, g_i)\mathbf{) + }K\ln(n/n_0)$$

where n_0 is the measure of informativeness in terms of how many observations the information in the parameter prior is based on. Kieseppä applies this idea to the AIC

and BIC results and shows that the informativeness of BIC parameter prior is equivalent to a sample with one observation [with $n_0 = 1$] whereas the parameter prior required in order to obtain the AIC result has the informativeness equivalent to $e^{-2}n$ observations. Hence the BIC parameter prior has constant informativeness independent of the number of observations contained in a given sample, whereas the AIC result is equivalent to the Bayesian result with the parameter prior which is more informative and its informativeness grows with the number of observations in the sample.

In fact there are infinitely many Bayesian models selection criteria – it all depends what value of n_0 one finds appropriate. Kieseppä (2001a) argues that this is a potential weakness of the Bayesian approach, because it seems to lose any normative character to the conclusions of model selection. In our view this flexibility is a positive attribute of Bayesian model selection methodology allowing one to reflect one's ideas about the way the parameters distributed within each individual model selection problem.

Finally, regarding Burnham and Anderson's contention in the previous section that imposition of a uniform model prior in BIC somehow commits us to the search for truth whereas their savvy model priors that lead to AIC do no such thing. In this section all of our results assume the uniform prior distribution over models. We have derived AIC under this assumption. Hence, their contention is incorrect.

4.5 Comparison between BIC and AIC

4.5.1 Statistical Consistency

Numerous sources [e.g., Keuzenkamp and McAleer (2001)] state that AIC is not a statistically consistent estimate. However, the BIC is statistically consistent.

Different questions can be asked about consistency of AIC.

1 '..[W]hether AIC is a consistent method of *maximizing* predictive accuracy in the sense of converging on the hypothesis with the greatest predictive accuracy in the large sample limit.

2 …[W]hether AIC is consistent estimator of predictive accuracy, which is a subtly different question from the first.

3 …[W]hether AIC converges to the smallest true *model* in a nested hierarchy of models.

The answer to the first two questions will be yes, …while the answer to the third is no, AIC is not consistent in this sense, but this fact does not limit its ability to achieve its goal.' [Forster (2001):113]

The AIC was designed as an estimator of predictive accuracy, so the charge should be that AIC fails to be consistent with respect to estimating the predictive accuracy. Forster shows that this is not the case. 'Akaike's own criterion minimizes the quantity $-2(\log L(\hat{\theta}_K) - K)$, which estimates $-2nA(\hat{\theta}_K)$. But note that this is a strange thing to estimate, since it depends on the number of seen data, *n*.' [Forster (1999):113] 'The correct response to the 'problem' is to divide the estimator and target by *n*, so that the target does not depend on the sample size. … AIC does provide a consistent estimate of predictive accuracy when it is properly defined.' [Forster (1999):114] It seems that Forster asserts that the AIC as it is commonly defined (see Introduction) is inconsistent with respect to predictive accuracy.

However, Kieseppä also discusses the question of consistency of AIC and reaches similar conclusions to Forster, but still uses the original form of AIC. So, it seems that either Forster is incorrect in saying that the proper definition of AIC score is the one divided by the number of data points in the sample, or Kieseppä is correct in using the original AIC.

Bandyopadhayay and Boik (1999) note that '[Forster's claim] is true in the special case of regression models where σ^2 is a known constant. In addition, if one is willing to assume that the approximating family is identical to the true family of models, then AIC is a consistent estimator of predictive accuracy. Forster's claim, however, is not true in general. If the approximating family misspecifies the true family, then AIC no longer is consistent.' [Bandyopadhayay and Boik (1999):S400]

Now, Forster turns to the charge that AIC is inconsistent with respect to estimating *K*. Forster considers the case of nested models, and distinguishes two cases. In the

first case, '…the true hypothesis will first appear in a model of dimension *K**, and in every model higher in the hierarchy.' [Forster (2001):114] Now the question arises of the desirability of estimating *k* as close as possible to *K**. Forster notes that in cases where data is drawn from quite a narrow range and supposing that we are choosing between LIN and PAR, '…for even quite large values of *n*, it may be best to select LIN over PAR, and better than any other family of polynomials higher in the hierarchy. Philosophically speaking, this is the interesting case in which a false model is better than a true model. However, for sufficiently high values of *n*, this will change, and PAR will be the better choice [because the problem of over fitting is then far smaller]. Again, this is an example in which asymptotic results are potentially misleading because they do not extend to intermediate data sizes.' [Forster (2001):114]

'In the second case the true hypothesis does not appear anywhere in the hierarchy of models. In this case the model bias will keep decreasing as we move up the hierarchy, and there will never be a point at which it stops decreasing. …There is no *universally valid* theorem that shows that BIC does better than AIC.' [Forster (2001):115] 'In both cases, the optimum model moves up the hierarchy as *n* increases. *In the first case, it reaches maximum value K*, and then stops.* The crucial point is that in all cases, the error of AIC (as an estimate of predictive accuracy) converges to zero as *n* tends to infinity.' [Forster (2001):115, italics added] Forster says that other information criteria are also consistent and he urges that it is most important what happens in the intermediate case and not in the limit.

It is rather difficult to see what exactly Forster claims at the end of the day. At the beginning of the section on consistency he seems to argue that AIC is not consistent with respect to estimating *k* and that this is of no consequence since this is not what AIC was designed to estimate anyhow, whereas the end of this section seems to suggest that AIC is actually consistent with respect to *K* {for example, 'After all, AIC does successfully converge on the true hypothesis!' [Forster (2001):115]}.

Kieseppä (2003) sheds clearer light on the issue by stating the result that '…when the sample size is large and the true curve is actually a horizontal straight line, the probability with which AIC will correctly recommend the model which contains

only horizontal straight lines is approximately 95%, and the probability that it will recommend the larger model which contains also all the other straight lines is approximately 5%.' [page 18] Unfortunately, Kieseppa had to omit the proof of this result due to the limitations of space. This result is in line with Forster's argumentation that AIC serves the purpose of picking hypotheses that are predictively accurate rather than that of finding/converging upon the true model with the minimum number of dimensions: '…it [AIC's recommendation] will with a very great probability be an acceptable choice, if the aim of the researcher is to find a curve which is "predictively accurate", although it will be a bad choice if her aim is to find out whether the true curve is a horizontal line or not.' [page 19]

4.5.2 Relative Performance

When the number of observations in a sample exceeds 8 [i.e., when $\ln(n) > 2$], BIC starts to give progressively greater weight to hypotheses with fewer adjustable parameters relative to AIC. Studies indicate that, all other things being equal, BIC performs better in set-ups where there are very few variables with strong effects whereas AIC performs best in contexts when there are several variables with moderate effects.

Let us return to the example that we used in chapter 3 to show how the AIC methodology works, and add the BIC to it. We repeat the table it here for convenience.

Cement hardening data with four regressor variables x_1, x_2, x_3 and x_4 and a response variable y

x_1	x_2	x_3	x_4	y
7	26	6	60	78.6
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7

1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

Where the regressor variables (in percentage of the weight) are: x1 = calcium aluminate (3CaO.Al2O3), x2 = tricalcium silicate (3CaO.SiO2), x3 = tetracalcium alumina ferrite (4CaO.Al2O3.Fe2O3) and x4 = dicalcium silicate (2CaO.SiO2); the response variable is y = total calories given off during hardening per gram of cement after 180 days.

Model	*K*	BIC (Δ)	AIC (Δ)	AICc (Δ)
12	4	0	0.45	0
124	5	2.73	0	3.13
123	5	2.65	0.04	3.16
14	4	3.46	3.77	3.32
134	5	3.4	0.75	3.88
234	5	8.31	5.6	8.73
1234	6	5.06	1.97	10.52
34	4	14.8	14.88	14.43
23	4	26.82	26.06	25.62
4	3	29.6	33.88	31.1
2	3	29.78	34.2	31.42
24	4	34.42	35.66	35.21
1	3	32.18	38.55	35.77
13	4	36.84	40.14	39.7
3	3	37.9	44.09	41.31

Ghosh and Samanta (2001):1143

As we can see, in this case BIC and AIC_c results broadly agree with one another.

5. Model Selection Methods and Scientific Realism

5.1 Introduction

In the previous chapters we surveyed the classical and some most recently developed approaches to model selection. We have seen that all methods have their strengths and weaknesses. We reviewed the objections to the methods and argued that none of them were devastating so long as one is aware of their foundations. In this chapter we explore what consequences, if any, the methods of model selection that we have considered have for some wider issues in the philosophy of science. In particular, what bearing these methods have on the debate on the scientific realism/anti-realism.

There are several versions of scientific realism available out there. However, we find that the following three theses capture the features of scientific realism well.

'Scientific Realism is a philosophical view about science that consists in three theses. *The Metaphysical Thesis*: the world has a definite and mind-independent structure. *The Semantic Thesis*: scientific theories should be taken at face value. They are truth-conditioned descriptions of their intended domain, both observable and unobservable. *The Epistemic Thesis*: mature and predictively successful scientific theories are well-confirmed and approximately true of the world.'
[Psillos (2007):226]

First, our focus is going to be the Epistemic Thesis. It presupposes both the metaphysical and the semantic theses. We will look at the epistemic thesis in its simplified form. That is, as the notion that predictively successful scientific theories are approximately true. There are numerous arguments both pro and con scientific realism in general and the epistemic thesis in particular. Arguably, among many arguments about scientific realism, the two most prominent ones so far are the no-miracles argument [some, including ourselves, consider it an intuition – cf. Worrall (1994)] and the argument from pessimistic meta-induction.

The no-miracles argument [NMA – this formulation is due to Putnam (1975)] purports to establish that predictive success of scientific theories licences the

inference to their [approximate] truth. That is, why else would a theory be predictively successful? It would be a 'miracle' if a theory were predictively successful but false.

In counterbalance to the NMA there is the argument from pessimistic meta-induction [cf. Laudan (1981)]. It has been noticed that there have been some very predictively successful theories in the history of science that eventually turned out to be, strictly speaking, false. The paradigm example is Newtonian mechanics, which was superseded by Einstein's theories of relativity. If such predictively successful theories like Newton's can be shown to be false, it may well be the case that other predictively successful theories that are currently entertained may eventually turn out to be false too.

Second, we shall look at an argument against the popular idea within the scientific realism field that science aims to find true theories.

5.2 Sober's Counterexamples

Model selection methods that we have been considering in this thesis attempt to capture predictive success, so it is natural to wonder whether these methods have any bearing on the issue of scientific realism. In fact, Elliot Sober (1999, 2002) takes up the challenge to show, firstly, contrary to the Epistemic Thesis of Scientific Realism, that there are false scientific theories [in our case probabilistic hypotheses] that are predictively successful. In fact Sober does not quite put it this way himself, but his argument clearly goes against the epistemic thesis. We consider this counterexample in section 5.2.1. The second counterexample purports to show that seeking truth and maximising predictive accuracy do not always go together. We consider this counterexample in section 5.2.2.

5.2.1 On the Epistemic Thesis of Scientific Realism

Let us consider Sober's example with differences in mean heights of corn plants in two different fields. Suppose that there are two adjacent fields in which corn grows, and suppose that we are interested in testing the following hypotheses about the average heights of wheat plants in both populations. We are doing so in the Neyman-Pearson way:

H_0: $|\mu_1 - \mu_2| = 0$

H_1: $|\mu_1 - \mu_2| \neq 0$

Sober argues that the null hypothesis is obviously false – surely the two population means cannot be *exactly* equal to several decimal places. Hence H_1 is obviously true. Scientists, however, routinely test such false hypotheses against true hypotheses. Assuming that scientists are rational and that the predictive accuracy is their only goal, Sober urges us to conclude that false hypotheses can be maximally predictively accurate – that is, sometimes even more predictively accurate than true hypotheses. Unless scientists believe that H_0 is more predictively successful than H_1, they would not bother testing such obviously false hypotheses against obviously true ones. Scientists seem to be willing to accept a false hypothesis as long as it is predictively successful. This argument, he contends, lends credence to methodological instrumentalism – 'the idea that theories are instruments for making predictions, [and] that predictive accuracy is the *only* consideration that matters [in science]' [Sober (1999): 4, 5].

There are counter arguments that deny that the sole goal of scientists is the accuracy of prediction. However, we have a different angle, and are willing to grant predictive accuracy as the goal in this particular example from statistics. Indeed in this thesis we have been looking at model selection exclusively for predictive purposes. We shall concentrate on the part of the argument that goes against the simplified version of the epistemic thesis of scientific realism. This argument against the epistemic thesis seems to be the following. The epistemic thesis asserts that predictively successful hypotheses are approximately true. Here we ostensibly have an example

of a hypothesis which is predictively successful [we assume so] but which is obviously false.

Recall from chapter 2 a feature of the NP hypothesis testing methodology is that it is important which hypothesis is H_0 and which is H_1. The method is more conservative towards H_0, so [depending on H_1] it can take quite a substantial difference between μ_1 and μ_2 to reject H_0 in favour of H_1 [the actual testing for this difference is done using respective sample means θ_1 and θ_2]. We contend that rather than thinking that a false hypothesis is more predictively accurate, scientists in this case use
H_0: $|\mu_1 - \mu_2| = 0$ as a place-holder for H_0: 'the difference between μ_1 and μ_2 is sufficiently small for us to disregard for our purposes'. What is 'sufficiently small' or statistically insignificant in the NP methodology is defined by the range of values of the parameter which is not the critical region [cf. section 2.2]. However, we briefly considered the notion of practical significance in section 2.1.2. We believe that our interpretation of what scientists take H_0 to stand for is consistent with this notion. There is also a notion of *substantive* significance [cf. Mayo and Spanos (2006) and references contained therein] that relies on the meta-statistical principle of *severity* of a statistical test. Unlike practical significance, the severity of a test and hence its substantive significance has a precise quantitative expression. It is calculated on the basis of observed data. The severity of a test is analogous to the concept of severe testing used by Popper (1968) – the more severe a test that a scientific theory survives, the more corroborated it is. Here the greater degree of severity confers more evidential support to a statistical hypothesis. Hence the concept of severity of a statistical test moves away from the behavioural interpretation of NP tests towards evidential support one. There is a lively debate on this subject of severity testing – cf. Achinstein (2003), Howson (1995, 1997), Mayo (1996, 2003, 2005). Here we shall no pursue this topic further, but note that it may constitute a fruitful avenue for further research in general, and in connection with Sober's views in particular.

Returning to the case that Sober discusses, it is important to note that in this case the testing of H_0 against H_1 at, say, the 5% level of significance is equivalent to finding a 95% confidence interval for $|\mu_1 - \mu_2|$ [cf. section 2.5]. In fact Sober seems to have found just such an interval by means of simulations: '[The] simulations closely agree

with the analytic solution that Branden Fitelson obtained, according to which [H_0] will be more predictively accurate (in expectation) than [H_1] precisely when
$| \mu_1 - \mu_2 | < 1.34898\ \sigma/\sqrt{n}$.' Sober (1999):21, footnote 7

Sober and Fitelson in fact found a confidence interval for $|\mu_1 - \mu_2|$! Once we reinterpret this case in such a way that rather than using a deliberately false hypothesis for greater predictive accuracy, the scientists implicitly check whether the differences in means fall within the confidence interval, i.e., they implicitly check whether $| \mu_1 - \mu_2 | < 1.34898\ \sigma/\sqrt{n}$, it is no longer obvious at all that the H_0 is false. We contend that scientists who use such point versus composite hypothesis tests simply do not spell out in detail what they intend to achieve by such testing, for it is often makes little practical difference for them if there is an insignificant deviation from zero. On this basis we argue that the putative connection between the falsity of a hypothesis and its predictive accuracy disappears. Scientists may be just a bit fast and loose with regards to describing the hypotheses, for the NP framework allows them to do so.

> 'If scientists interpret the [H_0] as saying that the means are no more than 2 inches apart, then they should *not* reject the [H_0] when they find that θ_1 and θ_2 differ by 1 inch in a large sample. However, this is precisely what they do. This argument generalizes to any setting of ε, large or small. The behaviour of scientists shows that they interpret [H_0] literally.'
> Sober (1999):28, notation modified to fit our usage

In our confidence interval for $| \mu_1 - \mu_2 |$ above, $\varepsilon = 1.34898\ \sigma/\sqrt{n}$. That is, ε is the critical value beyond which H_0 is rejected. As *n* increases, the critical value ε becomes smaller. In the limit as *n* tends to infinity, ε tends to zero. Hence, if this was the way that the NP method was used, in a large enough sample more or less *any* deviation in the difference between sample means θ_1 and θ_2 from zero would lead to rejection of H_0. We contend that in the quote above Sober's account of statistical practice is inaccurate. The users of NP tests often reduce the critical region [or, equivalently in our case of the confidence interval interpretation, they would increase the level of confidence beyond 95%] to account for the over-sensitivity of the test with large *n* to the tiniest differences in values – cf. our discussion of Lindley Paradox in section 2.2. The behaviour of scientists as we know it is consistent with our interpretation that they do not take H_0 literally.

So, if Sober's attempt to show that H_0 is obviously false does not succeed, is there a way to reformulate his counterexample? We think that there is, but it does not succeed either. Let us start with consideration of the notion of hypothesis 'acceptance' that Sober employs.

> 'In formulating the question as one about "acceptance", I leave open whether "acceptance" means *believing that the hypothesis is true* or *believing that it will be predictively accurate*. [footnote: Although I'll formulate the problem in terms of the concept of "acceptance", this is a matter of convenience; the dichotomous concept of acceptance could be replaced with the concept of degree of belief. Formulated in the latter way, the question would be whether the goal of science is to say how probable it is that various hypotheses are true, or to say how predictively accurate one should expect those hypotheses to be.]'
> Sober (1999):14

As it is usually understood, to accept a hypothesis within the NP framework means to behave as if it is true [cf. sections 2.2 and 2.3 for elaboration]. Thus in the example with corn plants accepting H_0 [when the difference between sample means falls within the confidence interval] involves behaving as if H_0 is true, and not behaving as if it was false, as Sober suggests. In the quote above Sober would be happy to replace this dichotomous concept of acceptance by talking of probability of hypotheses. As we know, there is no place for probabilities of hypotheses in the NP methodology. We would need to move to the Bayesian framework to use this concept sensibly. Let us try to recast the counterexample in a Bayesian way.

In a Bayesian rendition of the corn plant example the scientists would have to be explicit about what range of values they would expect the differences in the mean values of the two populations of plants to lie in. Let us then take the de facto confidence interval of a kind that Sober and Fitelson yielded in the NP example above as our null hypothesis and the interval outside the confidence interval as our alternative – thus null and alternative are exhaustive and mutually exclusive. For the sake of an argument let us suppose that $\varepsilon = 3$ so that our hypotheses are:

H_0: $(\mu_1 - \mu_2) \in [-3, 3]$

H_1: $(\mu_1 - \mu_2) \in (-\infty, -3) \cup (3, \infty)$

We need to assign prior probabilities to these hypotheses. Sober could argue that H_0 should be assigned much lower probability than H_1, possibly on the grounds that the interval of possible values that is suggested by H_0 is much shorter than that of H_1. So in this context rather than arguing that the null hypothesis is obviously false but is nonetheless deemed by the scientists predictively successful [as before], Sober could argue that the null has much lower probability of being true [prior to observing the difference in sample means], but it is still deemed more predictively successful than the alternative which has a higher probability. This argument sounds Popperian – recall our discussion of model priors in section 4.4.1. It could be argued in the spirit of Popper that H_0 has a much higher empirical content than H_1 – that is, there are many more possible observations that are incompatible with H_0 rather than with H_1. If this were the case then we would assign much lower prior probability to H_0 than to H_1 – in proportion to their respective empirical contents.

Unfortunately this argument does not work either. Choice of the interval [-3, 3] may suggest that the scientists have an expectation of $(\mu_1 - \mu_2)$ to lie within this interval, presumably on the basis of their background knowledge. This suggests that at the very least there is no reason to set the prior on the null much lower than that of the alternative.

5.2.2 Truth and Predictive Accuracy

In philosophy of science it is commonly thought that in addition to the three theses cited in section 5.1, scientific realist is committed to seeking truth as the aim of science. How does it connect with the aim of predictive accuracy, which we have been assuming in this thesis? The two aims seem to occur together – we would expect true theories to be most predictively accurate [cf. Nagel (1979):139]. However, Sober (1999) uses the following example to show that seeking truth and maximising predictive accuracy do not always coincide.

> ‘Suppose that one of the buses numbered 1-10 takes you right to Fred’s door, while the other nine take you very far away; on the other hand, all of the buses numbered 11-20 go very near Fred’s house, though none of them goes right to his door. … If your goal is to get as close as possible to Fred’s house, you should take a bus numbered 11-20. The point is this: even if a bus with a

> low number is the one that goes closest to Fred's house, it doesn't follow that the best way to get close to Fred's house is to take a low-numbered bus. … This suggests that there may be inference problems in which trying to find the truth and trying to maximise predictive accuracy lead to different decisions. The bus example suggests that this may be possible even if no hypothesis is more predictively accurate than the truth.'
> Sober (1999):13

Here finding the truth maximises predictive accuracy, but the probability of picking the true hypothesis is low whereas the alternative is to pick a hypothesis which is very close to truth with certainty.

We agree with Sober that in his bus to Fred's house example trying to find the truth and trying to maximise predictive accuracy leads to different decisions, and that we would also take a bus numbered 11-20. However, we argue that if we refine the goal of finding the truth in a quite natural way, then we restore the connection between truth and predictive accuracy.

In this example there is uncertainty as to how far a given bus would take us from Fred's house. We suggest that this uncertainty can be handled probabilistically. In this case we can substitute the goal of seeking the truth by the goal of minimising the expectation [in the statistical sense of a probability weighted average – cf. chapter 1] of the divergence from truth. Then trying to minimise the expectation of the divergence from truth and trying to maximise predictive accuracy lead to the same decision – choosing a higher numbered bus. In contexts of uncertainty of the kind that is there in the bus to Fred's house example a scientific realist should refine her aim from seeking truth to minimising expected divergence from truth. Indeed, it is not accidental that scientific realists use the concept of *approximate* truth in the epistemic thesis of scientific realism. Likewise the aim of approximate truth is more realistic than that of truth *simpliciter*. Minimising expected divergence from truth can be thought of as operationalising the concept of approximate truth.

Using the bus to Fred's house example Sober argues against the principle that:

(*) If you want to maximize A and T maximizes A, then the best way to maximize A is to try to maximize T. [Sober (1999):12]

We think that once we refine our goal in the way suggested, the bus to Fred's house example provides support to the principle (*). In the example we are urged to choose a bus 11-20, because such a choice would minimise the distance to the Fred's house [which is the proxy for actual truth in the example]. Arguably such minimisation of distance to truth can be thought of as maximisation of truth. Indeed, this is the very idea behind the AIC framework, which aims to minimise the K-L divergence to 'truth'. In this example the expected distance to Fred's house in each case is a probability-weighted average of minimum Euclidean distances within Fred's house that each bus from No. 1 to 10 and from No. 11 to 20 respectively brings one. For instance, using the Principle of Indifference as per section 4.1.2.1 the probability of picking the bus to 'truth' is 0.1. It is obvious that the expected average distance would be shorter if one were to choose a bus from No. 11 to 20. This does not violate the idea that the search for the minimum expected divergence [in this case Euclidean distance] from truth and search for predictively accuracy go hand in hand.

5.3 AIC, BIC and the Epistemic Thesis of Scientific Realism

In section 5.2.1 we argued contra Sober that when properly understood the Neyman-Pearson methodology was logically consistent with the simplified epistemic thesis of scientific realism [that predictively successful scientific theories are approximately true]. There we attempted to give Sober's corn plants example a Bayesian twist, but conclude that it was not successful either.

Let us now see what relation if any the AIC and BIC methods have to the epistemic thesis of scientific realism.

In section 3.2.2 we saw that the AIC was derived as the asymptotically unbiased estimator of relative expected Kullback-Leiber divergence from the putative 'truth'. In section 4.3.2 we saw that the BIC can also be thought as estimating the Kullback-Leibler divergence from the 'truth', but in a Bayesian way and on the assumption that the predictive performance of models in the choice set is properly judged by the logarithmic scoring rule. There we referred to the AIC or BIC-best model as quasi-true in the precise sense that such a model is relatively K-L closer to the 'truth' than any other model within the choice set, although the quasi-true model can still be

arbitrarily far away from such 'truth' – we have no idea about the absolute rather than relative divergence. In section 4.4 we showed the connection between AIC and BIC within the Bayesian setting. Therefore, the following deliberations apply to both AIC and BIC methods, although for ease of presentation we will be mentioning AIC only.

So in the AIC model selection we set out to find a quasi-true model within the range of models that we think may be relevant to the problem at hand. Suppose that the AIC-best model that we have found actually turns out to be predictively successful. Does it then mean that this model is approximately true? Unfortunately the answer has got to be – not necessarily. Just because the AIC method was explicitly set up to approximate the relative K-L divergence from truth in a given set of models in order to maximise predictive accuracy, and the AIC-best model is then found to be predictively successful, this is no argument for success in approximating the truth. We simply do not know whether we succeeded in this endeavour – to re-iterate, it is still possible for the AIC-best model to be arbitrarily far away from the truth, nothing in the AIC method precludes this. The person in the bus to Fred's house example [let us call her Daisy] in section 5.2.2 potentially has epistemic access to how approximately true her selected hypothesis turns out to be. All she has to do is to bring a tape measure [or some device that utilises the Global Positioning System – we are going to take it for granted that there is some reliable method of measuring the distance that Daisy can use] with her and measure the actual distance from the bus stop at which she eventually gets off to Fred's house. Notice that this measure simultaneously serves as a measure of the predictive success of the hypothesis that Daisy selected and as a measure of its approximate truth [or divergence from truth, which we use interchangeably]. In our case we also have epistemic access to how predictively successful our AIC-best model has turned out to be, but crucially we do not have the luxury of epistemic access to the actual divergence between our AIC-best model and truth. Hence the AIC model selection methodology does not yield an argument against the epistemic thesis of scientific realism either. That is, there is no way to argue that, despite our best efforts to the contrary, our predictively successful AIC-best model is in fact further away from the truth than all the other models within the choice set. In other words there is no way to show by means of an argument that our predictively successful model is in fact quasi-false.

Of course it is tempting to argue that it is highly unlikely that the predictive success of our AIC-best model is attributable to anything else except for its relative closeness to truth. However, that puts us back on the familiar grounds of the No-Miracles Argument, which in its turn familiarly counted by a type of pessimistic meta-induction [which we contend is more accurately referred to as 'the pessimistic induction from the history of science' – cf. Godfrey-Smith (2003):177]. There are instances of model selection not leading to predictively successful models or yielding models that are predictively successful for a while, and then cease to be such, particularly in a field such as economics where successful predictive modelling is notoriously elusive.

We thus conclude that the model selection methodologies considered in this thesis are neutral with respect to the arguments regarding the epistemic thesis of scientific realism. They do, however, serve the purpose of recasting the familiar arguments in the new light, which can be illuminating.

There is a further worry that our neutrality conclusion could play into the hands of the antirealists since being a scientific realist is not required in order to understand the model selection methods considered in this thesis. This worry seems to stem from an argument that antirealists such as van Fraassen (1980) put forward, viz., the argument that scientific realism is unnecessarily inflationist. That is, the statement that predictively successful scientific theories are [approximately] true is logically stronger than the statement that predictively successful scientific theories are empirically adequate. One can maintain the latter [as van Fraassen does] while remaining agnostic about the former, and not lose anything scientifically important in the process.

We think that there is no onus on someone who finds the No-Miracles Argument plausible, and accepts the philosophical position of scientific realism, to provide further justification of their *philosophical* stance by having to demonstrate what useful purpose their commitment to scientific realism serves in a particular field of science. Naturally, it is superb when one's philosophical views lead to advances in the empirical realm, but it would in our view be too strong a requirement for

judgement of viability of such views. In this work there is nothing to undermine the plausibility intuitions behind the No-Miracles Argument. We venture that our neutrality conclusion *really is neutral* with regards to the debate between scientific realists and antirealists.

Finally, one may wonder, as indeed some have done, whether the model selection methodologies that we consider in this thesis can be used to rationally reconstruct the key moments of model choice in history of science. For example, Forster and Sober (1994):14-15 argue that the AIC methodology provides a reason for choice of Copernicus's astronomy as compared to Ptolemy's astronomy. Kieseppä (1997):37-39 points out that the AIC framework has not been proved to apply to periodic functions [in fact there are examples of failures of such applications] and that neither of the astronomical systems are in the form of statistical hypotheses specifying different probability distributions for the observable quantities. On these grounds we agree with Kieseppä [ibid.] that reconstructing this case in terms of AIC model selection is implausible. We struggle to come up with another case in history of science which could be reconstructed in the model selection fashion with some plausibility. Does this affect our analysis of the relation between the model selection criteria and the issue of scientific realism? We believe that it does not. It does, however, remind us of exactly which types of models the model selection methods are applicable to.

6. Conclusion

In this thesis we have considered the classical approaches such as those due to Ronald Fisher and to Jerzy Neyman and Egon Pearson, as well as more recent approaches of Akaike Information Criterion and Bayes Information Criterion to the problem of model selection for predictive purposes. We find that the Fisherian approach can be thought of as an approach to the problem of model selection only in a rather Pickwickian sense, the Neyman-Pearson method in a limited but nonetheless viable sense, and the AIC and BIC methods in the fully-blown sense of aiming to choose a model with the optimal mathematical structure. We then move onto considering the numerous objections that have been raised in the recent philosophical literature to the AIC and BIC methods. Chief among these objections is the Subfamily Problem [about rendering the method defunct by fixing of adjustable parameters within models in the choice set in the light of the sample of data at hand] that we look into within the AIC setting, and the issues with the nesting of models and the ostensible requirement for inclusion of the 'true' model within the choice set for the BIC method. Upon careful consideration of the foundations of the AIC and BIC and of the arguments involved, we argue that at the very least none of these issues are devastating for the two methodologies of model selection. We then show that there are ways to connect AIC and BIC within the setting of the Bayesian theory of statistics and argue *pace* Burnham and Anderson (2004) that the way in which one sets model priors does not imply any particular attitude towards the aim of using the BIC method. We also show that within the Bayesian setting there are in fact infinitely many model selection criteria that have similar form to AIC and BIC. Namely, they penalise the maximum likelihood of the best-fitting element of the given model by a function of the variance of the parameter prior multiplied by the number of adjustable parameters that the model contains. We argue that this state of affairs is favourable for the scientists who can choose the prior according to their ideas and the background knowledge about the problem at hand – the diagrams in section 4.1.2.2 exhibit the amazing flexibility of priors. We then provide an overview of the circumstances under which the AIC and BIC are said to perform better than one another.

Then we consider two counterexamples that are due to Elliott Sober (1999 and 2002). The counterexamples were against the simplified form of the Epistemic Thesis of Scientific Realism [that predictively successful theories are approximately true] and against the idea popular among scientific realists that the aim of science is to search for theories that are true. In the former counterexample Sober argues that the way that scientists put the Neyman-Pearson methodology to in every day use indicates that they expect to be predictively accurate hypotheses that are obviously false. Assuming that the scientists are rational, Sober concludes that scientists' behaviour implies that they are methodological instrumentalists. Sober states that methodological instrumentalism commits one to using theories as tools for making predictions, and to having predictive accuracy as the one and only goal of scientific endeavour. We go along with Sober's assumption about the goal of predictive accuracy, but argue that he misconstrues the way that scientists use the NP methodology. Contrary to Sober, we argue that the Neyman-Pearson methodology is logically consistent with the epistemic thesis of scientific realism. We attempt to give Sober's counterexample Bayesian rendition using some ideas in the spirit of Karl Popper, but conclude that his argument does not succeed this way either.

Sober's second counterexample attempts to show that searches for theories that are true and for the theories that are predictively successful do not always coincide. Sober thus argues that the popular idea in the scientific realist camp that the goal of scientific enterprise is to find theories that are true can go against maximising predictive accuracy, the latter arguably being a desirable feature of any scientific theory. We argue that his counterexample does not succeed in demonstrating that the link between [at least approximate] truth and predictive accuracy is bogus.

However, we go on to argue that the AIC and BIC methods are actually neutral with regards to the debate about the epistemic thesis of scientific realism. That is, these methods neither lend support to nor go against the epistemic thesis. On the other hand, we think that AIC and BIC do provide a different angle from which to view the familiar arguments within scientific realism, namely, the No-Miracles Argument and the Pessimistic Meta-Induction. Our view is that the conclusion of neutrality of our formal methods of model selection with respect to some issues within scientific realism is indicative of the general idea that it is extremely rare for purely formal

methods to settle philosophical disputes. Nonetheless, in trying to do so one at the very least gains additional valuable insights.

It is important for us to emphasise that notwithstanding the fact that in the domain of the AIC and BIC methods the talk of simplicity and its predictive accuracy maximising virtue has been pervasive, in this thesis simplicity is hardly mentioned, and when it is mentioned, it is only as short hand for 'relatively fewer number of adjustable parameters'. The reason for not paying homage to simplicity in the AIC and BIC context is this. The AIC was designed to provide unbiased estimates of relative expected Kullback-Leibler divergence from a set of models to the 'truth', where the penalty for complexity in the form of the number of adjustable parameters arose as a by-product in order to correct the asymptotic bias. In the BIC the penalty for complexity arose as a by-product of approximating often computationally intractable integrated likelihoods. So in neither of these frameworks was simplicity built-in as an important consideration. Moreover, we find that we do not lose anything by ignoring simplicity and treating it as an epiphenomenon.

Finally we suggest that the best handle on the problem of model selection is to be gained by applying different approaches to the same issue with full awareness of the foundational and philosophical issues involved. We sincerely hope to have at least partially served this purpose in this thesis.

7. References

Achinstein, P. (2003) 'The Book of Evidence', OUP: Oxford, UK

Agresti, A. and Finlay, B. (2009) (4th ed.) "Statistical Methods for the Social Sciences", New Jersey, USA: Pearson Prentice Hall

Aitkin, M. (1991) 'Posterior Bayes Factors', *Journal of the Royal Statistical Society* Series B **53**: 111-142

Akaike, H (1973) 'Information Theory and an Extension of the Maximum Likelihood Principle', in B. N. Petrov and F. Csáki (eds.), 1973, 2nd International Symposium on Information Theory. Budapest: Akadémiai Kiadó, 267-81, reprinted in Johnson, N. L. and Kotz, S. (eds.) (1991) 'Breakthroughs in Statistics', *Vol. 1: Foundations and Basis Theory*, New York: Springer-Verlag, pp. 599-624

Akaike, H. (1974) 'A New Look at the Statistical Model Identification', *IEEE Transactions on Automatic Control*, **19**: 716-723

Akaike, H. (1985) 'Prediction and Entropy', in Atkinson, C. A. and Fienberg, S. E. (eds.), *A Celebration of Statistics*, Springer-Verlag: New York: 1-23

Aldrich, J. (2005) 'The Statistical Education of Harold Jeffreys', *International Statistical Review* **73**, Vol. 2: 289-307

Bandyopadhayay, P., Boik, R, Basu, P. (1996) 'The Curve-Fitting problem: A Bayesian Approach', *Philosophy of Science*, **63**, Supplement: Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association. Part I: Contributed Papers: S264-S272

Bandyopadhayay, P., Boik, R. (1999) 'The Curve-Fitting problem: A Bayesian Rejoinder', *Philosophy of Science*, **66**, Supplement. Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association. Part I: Contributed Papers: S390-S402

Baron, J. (2008) 'Thinking and Deciding' (4th ed.), CUP: Cambridge, UK

Bozdogan, H. (1987) 'Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions', *Psychometrica* **52**: 345-370

Bozdogan, H. (2000) 'Akaike's Information Criterion and Recent Developments in Information Complexity', *Journal of Mathematical Psychology* **44**: 62-91

Bunge, M. (1963) 'The Myth of Simplicity: Problems of Scientific Philosophy', Prentice-Hall Inc.: Englewood Cliffs, N.J., USA

Burnham, K. P. and Anderson, D. R. (2002) 'Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach' (2nd ed.), Springer-Verlag: New York, USA

Burnham, K. P. and Anderson, D. R. (2004) 'Understanding AIC and BIC in Model Selection', *Sociological Methods & Research* **33**, No. 2: 261-304

Busemeyer, J. and Wang, Y. (2000) 'Model Comparisons and Model Selections Based on Generalization Criterion Methodology', *Journal of Mathematical Psychology* **44**: 171-189

Cavanaugh, J. E. and Neath, A. A. (1999) 'Generalizing the Derivation of the Schwarz Information Criterion', *Communications in Statistics – Theory and Methods* **28**: 49-66

Chakrabarti, A. and Ghosh, J. K. (2011) 'AIC, BIC and Recent Advances in Model Selection', in Gabbay, D. V., Bandyopadhyay, P. S., Forster, M. R., Thagard, P. and Wood, J. *Philosophy of Statistics*, Elsevier B. V.: Holland: 583-605

Crosby, A. W. (2002) 'Throwing Fire: Projectile Technology Through History', CUP: Cambridge, UK

Dawid, P. and Senn, S. (2011) 'Statistical Model Selection', in Christie, M., Cliffe, A., Dawid, P. and Senn, S. (eds.) *Simplicity, Complexity and Modelling*, John Wiley & Sons, Ltd.: Chichester, West Sussex, UK: 11-33

De Vito, S. (1997) 'A Gruesome Problem for the Curve-Fitting Solution', *The British Journal for the Philosophy of Science* **48**, No. 3: 391-396

Dowe, D. L., Gardner, S. and Oppy, G. (2007) 'Bayes not Bust! Why Simplicity is no Problem for Bayesians', *The British Journal for the Philosophy of Science* **58**, No. 4: 709-754

Earman, J. (1986) 'A Primer on Determinism', *University of Western Ontario Series in Philosophy of Science* **32**, D. Reidel Publishing Company: Dordrecht, the Netherlands

Ehm, W. and Gneiting, T. (2009, Addendum 2010) 'Local Proper Scoring Rules', University of Washington, Department of Statistics, Technical Report No. 551

Efron, B. (1986) 'Why Isn't Everyone a Bayesian?', *The American Statistician* **40**: 1-5

Findley, D. and Parzen, E. (1995) 'A Conversation with Hirotugu Akaike', *Statistical Science* **10**: 104-117

Fisher, R. A. (1922) 'On the Mathematical Foundations of Theoretical Statistics', *Philosophical Transactions of the Royal Society of London (A)* **222**: 309-368

Fisher, R. A. (1956) 'Statistical Methods and Statistical Inference', Oliver and Boyd: Edinburgh, UK

Fitelson, B. (1999) 'The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity', *Philosophy of Science* **66**, Supplement: Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association. Part I: Contributed Papers: S362-S378

Forster, M. R. (1995a) 'Bayes and Bust: Simplicity as a Problem for a Probabilist Approach to Confirmation', *The British Journal for the Philosophy of Science* **46**: 399-424

Forster, M. R. (1995b) 'The Golfers' Dilemma: A Reply to Kukla on Curve-Fitting', *The British Journal for the Philosophy of Science* **46**, No. 3: 348-360

Forster, M. R. (1999) 'Model Selection in Science: The Problem of Language Variance', *The British Journal for the Philosophy of Science* **50**, No. 1: 83-102

Forster, M. R. (2000) 'Key Concepts in Model Selection: Performance and Generalizability', *Journal of Mathematical Psychology* **44**: 205-231

Forster, M. R. (2001) 'The New Science of Simplicity', in Zellner, A., Keuzenkamp, H. and McAleer, M. (eds.) *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*. CUP: Cambridge, UK: 83-119

Forster, M. R. (2002) 'Predictive Accuracy as an Achievable Goal of Science', *Philosophy of Science* **69**, No. S3: S124-S134

Forster, M. (2006) 'A Philosopher's Guide to Empirical Success', in *Philosophy of Science Assoc. 20th Biennial Meeting (Vancouver): PSA 2006 Contributed Papers*

Forster, M. R. and Sober, E. (1994): 'How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions', *The British Journal for the Philosophy of Science* **45**, No. 1: 1-35

Forster, M. and Sober, E. (2004): 'Reply to Boik and Kruse,' in Mark Taper and Subhash Lele (eds.), *Likelihood and Evidence*, Chicago and London: University of Chicago Press: 181-190

Frigg, R. and Hartmann, S. (2006) 'Models in Science', *Stanford Encyclopedia of Philosophy*, URL: http://plato.stanford.edu/entries/models-science/

Gibbons, J. D. and Chakraborti, S. (2003) 'Nonparametric Statistical Inference' (4th ed. revised and expanded), Marcel Dekker, Inc: New York, USA

Gillies, D. (1973) 'An Objective Theory of Probability', Methuen & Co Ltd: London, UK

Gillies, D. (2000) 'Philosophical Theories of Probability', Routledge: London, UK

Ghosh, J. K. and Samanta, T. (2001) 'Model Selection – an Overview', *Current Science* **80**, No. 9: 1135-1144

Gneiting, T. and Raftery, A. E. (2007) 'Strictly Proper Scoring Rules, Prediction, and Estimation', *Journal of the American Statistical Association* **102**, No. 477: 359-378

Godfrey-Smith, P. (2003) 'Theory and Reality', The University of Chicago Press: Chicago, USA

Good, I. J. (1950) 'Probability and the Weighing of Evidence', Griffin: London, UK

Good, I. J. (1952) 'Rational Decisions', *Journal of the Royal Statistical Society, Series B (Methodological)* **14**, No. 1: 107-114

Hájek, A. (2009) 'Interpretations of Probability', *Stanford Encyclopedia of Philosophy*, URL: http://plato.stanford.edu/entries/probability-interpret/

Hitchcock, C. and Sober, E. (2004) 'Prediction vs. Accommodation and the Risk of Overfitting', *The British Journal for the Philosophy of Science* **55**: 1-34
Hoeting, J., Madigan D., et *al.* (1999) 'Bayesian Model Averaging: A Tutorial (with comments)', *Statistical Science* **14**: 382-417

Hoover, K. D. and Siegler, M. V. (2008) 'Sound and Fury: McCloskey and Significance Testing in Economics', *Journal of Economic Methodology* **15**, No.1: 1-37

Howson, C. (1987) 'Popper, Prior Probabilities, and Inductive Inference', *The British Journal for the Philosophy of Science* **38**, No. 2: 207-224

Howson, C. (1988) 'On the Consistency of Jeffrey's Simplicity Postulate, and Its Role in Bayesian Inference', *The Philosophical Quarterly* **38**, No. 150: 68-83

Howson, C. (1995) 'Theories of Probability', *The British Journal for the Philosophy of Science* **46**: 1-32

Howson, C. (1997) 'A Logic of Induction', *Philosophy of Science* **64**: 268-290

Howson, C. (2007) 'Logic with Numbers', *Synthese* **156**: 491-512

Howson, C. (2008) 'De Finneti, Countable Additivity, Consistency and Coherence', *The British Journal for the Philosophy of Science* **59**: 1-23

Howson, C. and Urbach, P. (2006) 'Scientific Reasoning: The Bayesian Approach' (3rd ed.), Open Court: Chicago and La Salle, Illinois, USA

Hurvich, C. M. and Tsai, C.-L. (1989) 'Regression and Time Series Model Selection in Small Samples', *Biometrika* **76**: 297-307

Jaynes, E. T. (1957) 'Information Theory and Statistical Mechanics', *The Physical Review* **106**, No. 4: 620-630

Jeffreys, H. (1961) 'Theory of Probability' (3rd ed.), OUP: Oxford, UK

Jeffreys, H. (1973) 'Scientific Inference', (3rd ed.) Cambridge: CUP

Kass, R. E. and Raftery, A. E. (1995) 'Bayes Factors', *Journal of the American Statistical Association* **90**, No. 430: 773-795

Kass, R. E. and Wasserman, L. (1995) 'A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion', *Journal of the American Statistical Association* **90**: 928-934

Keuzenkamp, H. and McAleer (1995) 'Simplicity, Scientific Inference and Econometric Modelling', *The Economic Journal* **105**: 1-21

Keynes, J. M. (1921) 'A Treatise on Probability', McMillan & Co: London, UK

Kieseppä, I. A. (1997) 'Akaike Information Criterion, Curve-Fitting, and the Philosophical Problem of Simplicity', *The British Journal for the Philosophy of Science* **48**: 21-48

Kieseppä, I. A. (2001a) 'Statistical Model Selection Criteria and Bayesianism', *Philosophy of Science*, **68**, Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers: S141-S152

Kieseppä, I. A. (2001b) 'Statistical Model Selection Criteria and the Philosophical Problem of Underdetermination', *The British Journal for the Philosophy of Science* **52**: 761-794

Kieseppä, I. A. (2003) 'AIC and Large Samples', *Philosophy of Science* **70**, No. 5: 1265-1276

Kolmogorov, A. N. (1956) 'Foundations of Probability', (2nd English ed.), Chelsea Publishing Company: New York, USA

Konishi, S. and Kitagawa, G. (1996) 'Generalised Information Criteria in Model Selection', *Biometrika* **83**: 875-890

Konishi, S. and Kitagawa, G. (2008) 'Information Criteria and Statistical Modeling', *Springer Series in Statistics*, Springer Science + Business Media, LLC: New York, USA

Kuha, J. (2004) 'AIC and BIC – Comparisons of Assumptions and Performance', *Sociological Methods & Research* **33**, No. 2: 188-229

Kuhn, T. (1977) 'Objectivity, Value Judgment, and Theory Choice' in *The Essential Tension: Selected Studies in Scientific Tradition and Change*, The University of Chicago Press: Chicago, USA: 320-339

Kukla, A. (1995) 'Forster and Sober on the Curve-Fitting Problem', *The British Journal for the Philosophy of Science* **46**, No. 2: 248-252

Kullback, S. and Leibler, R. A. (1951) 'On Information and Sufficiency', *The Annals of Mathematical Statistics* **22**, No. 1: 79-86

Laudan, L. (1981) 'A Confutation of Convergent Realism', Philosophy of Science **48**, No. 1: 19-49

Lehmann, E. L. (1986) 'Testing Statistical Hypotheses', (2nd ed.), *Wiley Series in Probability and Mathematical Statistics*, John Wiley & Sons: New York, USA

Lehmann, E. L. (1990) 'Model Specification: The Views of Fisher and Neyman, and Later Developments', *Statistical Science* **5**, No. 2: 160-168

Lehmann, E. L. (1993) 'The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?', *Journal of the American Statistical Association* **88**, No. 424: 1242-1249

Lenhard, J. (2006) 'Models and Statistical Inference: The Controversy between Fisher and Neyman-Pearson', *The British Journal for the Philosophy of Science* **57**: 69-91

Lindley, D. (2000) 'The Philosophy of Statistics (with comments)', *The Statistician* **49**: 293-337

Madigan, D. and Raftery, A. E. (1994) 'Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window', *Journal of the American Statistical Association* **89**, No. 428: 1535-1546

Mayo, D. G. (1996) 'Error and the Growth of Experimental Knowledge', The University of Chicago Press: Chicago, USA

Mayo, D. G. (2003) 'Could Fisher, Jeffreys and Neyman Have Agreed? Commentary on J. Berger's Fisher Address', *Statistical Science* **18**: 19-24

Mayo, D. G. (2005) 'Philosophy of Statistics', in Sarkar, S. and Pfeifer, J. (eds.) *Philosophy of Science: An Encyclopaedia*, Routledge: London, UK

Mayo, D. G. and Spanos, A. (2006) 'Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction', *The British Journal for the Philosophy of Science* **57**: 323-357

Mikkelson, G. M. (2006) 'Realism versus Instrumentalism in a New Statistical Framework', *Philosophy of Science* **73**: 440-447

Nagel, E. (1979) 'The Structure of Science', Hackett Publishing: Indianapolis, USA

Newbold, P. (1995) 'Statistics for Business and Economics', (4th ed.) Prentice Hall International (UK) Ltd: London, UK

Popper, K. R. (1968) 'The Logic of Scientific Discovery' (revised ed.), Hutchinson & Co (Publishers) Ltd.: London, UK

Popper, K. R. (1982) 'The Open Universe: An Argument for Indeterminism', Routledge: London, UK

Psillos, S. (1999) 'Scientific Realism: How Science Tracks Truth', Routledge: London, UK

Psillos, S. (2007) 'Philosophy of Science A – Z', Edinburgh University Press Ltd.: Edinburgh, UK

Putnam, H. (1975) 'Philosophical Papers Vol. 1, Mathematics, Matter and Method', CUP: Cambridge, UK

Raftery, A. E. (1995) 'Bayesian Model Selection in Social Research (with discussion)', *Sociological Methodology* **25**: 111-163

Royall, R. M. (1997) 'Statistical Evidence: A Likelihood Paradigm', Chapman & Hall/CRC: New York, USA

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986) 'Akaike Information Statistics', KTK Scientific Publisher/D. Reidel Publishing: Tokyo/Dordrecht

Schwarz, G. (1978) 'Estimating the Dimension of a Model', *Annals of Statistics* **6**: 461-465

Shannon, C. E. and Weaver, W. (1949) 'The Mathematical Theory of Communication', The University of Illinois Press: Chicago, USA

Silvey, S. D. (1975) 'Statistical Inference', *Monographs on Applied Probability and Statistics*, Chapman and Hall: London, UK

Sober, E. (1990) 'Contrastive Empiricism', in Savage, W. (ed) 'Studies in the Philosophy of Science' Minneapolis: University of Minnesota Press, *Scientific Theories* **14**: 392-412

Sober, E. (1996) 'Parsimony and Predictive Equivalence', *Erkenntnis* **44**: 167-197

Sober, E. (1999) 'Instrumentalism Revisited', *CRITICA, Revista Hispanoamericana de Filosofia* **31**: 3-39

Sober, E. (2001) 'What is The Problem of Simplicity?', in Zellner, A., Keuzenkamp, H. and McAleer, M. (eds.) *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*. Cambridge: Cambridge University Press: 13-32

Sober, E. (2002): "Instrumentalism, Parsimony, and the Akaike Framework." PSA 2000 – Proceedings of the Philosophy of Science Association **69**: S112-S123

Sober, E. (2008) 'Empiricism', in Psillos, S. and Curd, M. (eds.) 'The Routledge Companion to Philosophy of Science', Routledge

Spanos, A. (2001) 'Parametric vs Non-Parametric Inference: Statistical Models and Simplicity', in Zellner, A., Keuzenkamp, H. A. and McAleer, M. (eds.) *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*, CUP: Cambridge, UK: 181-206

Spiegelhalter, D. J., Best N. G., Carlin, B. P. and van der Linde, A. (2002) 'Bayesian Measures of Model Complexity and Fit (with discussion)', *Journal of Royal Statistical Society* B **64**: 583-639

Stewart, I. (2002) 'Does God Play Dice? The New Mathematics of Chaos' (2nd ed.), Blackwell: Malden, Massachusetts, USA

Stuart, A. (1962) 'Basic Ideas of Scientific Sampling', Charles Griffin & Co. Ltd.: High Wycombe, UK

Stuart, A. (1984) 'The Ideas of Sampling', Charles Griffin & Co. Ltd.: High Wycombe, UK

Sugiura, N. (1978) 'Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections', *Communications in Statistics: Theory and Methods* **7**: 13-26

Urbach, P. (1989) 'Random Sampling and the Principle of Estimation', *Proceedings of the Aristotelian Society*, New Series **89**: 143-164

Van Fraassen. B. (1980) 'The Scientific Image', Clarendon Press: Oxford, UK

Wasserman, L. (2000) 'Bayesian Model Selection and Model Averaging', *Journal of Mathematical Psychology* **44**: 92-107

Wagenmakers, E. and Farrell, S. (2004) 'AIC Model Selection Using Akaike Weights', *Psychonomic Bulletin & Review* **11**: 192-196

Weakliem, D. (1999) 'A Critique of the Bayesian Information Criterion for Model Selection (with comments and discussion)', *Sociological Methods and Research* **27** The Special Issue on the Bayesian Information Criterion: 359-443

Williamson, J. (1999) 'Countable Additivity and Subjective Probability', *The British Journal for the Philosophy of Science* **50**: 401-416

Williamson, J. (2007) 'Motivating Objective Bayesianism: from Empirical Constraints to Objective Probabilities', in Harper, W. L. and Wheeler, G. R. (eds.): *Probability and Inference: Essays in Honour of Henry E. Kyburg Jr*. College Publications: London, UK: 155-183

Williamson, J. (2010) 'In Defence of Objective Bayesianism', OUP: Oxford, UK

Worrall, J. (ed.) (1994) 'The Ontology of Science', Dartmouth: Aldershot, UK

Worrall, J. (2002) ‘New Evidence for Old’, in Gardenfors, P., Wolenski, J., Kijania-Placek, K. (eds.) *In the Scope of Logic, Methodology and Philosophy of Science*, Kluwer Academic Publishers: Dordrecht: 191-209

Zahar, E. (1973) ‘Why did Einstein’s Programme supersede Lorentz’s? (I)’, *The British Journal of the Philosophy of Science* **24**, No. 2: 95-123

Zellner, A., Keuzenkamp, H. A. and McAleer, M. (eds.) (2001) ‘Simplicity, Inference and Modeling: Keeping It Sophisticatedly Simple’, CUP: Cambridge, UK