4

Quantifying the Evidence About Science Hypotheses

Richard Arthur Leibler (1914–2003) was born in Chicago, Illinois on March 18, 1914. He received a Bachelors and Masters degree in mathematics from Northwestern University and a Ph.D. in mathematics at the University of Illinois (1939). After serving in the Navy during the war, he was a member of the Institute for Advanced Study at Princeton and a member of the von Neumann Computer Project 1946–1948. From 1948–1980 he worked for the National Security Agency (1948–1958 and 1977–1980) and the Communications Research Division of the Institute for Defense Analysis (1958–1977). He then was the president of Data Handling Inc., a consulting firm for the Intelligence Community. He received many awards, including the Exceptional Civilian Service Award.

The ability to simply rank science hypotheses and their models is a major advance over what can be done using null hypothesis tests. However, much more can be done, all under the framework of "strength of evidence," for hypotheses in the *a priori* candidate set. Such evidence is exactly what Platt (1964) wanted in his well-known paper on *strong inference*. I begin by describing four new evidential quantities.

4.1 Δ_i Values and Ranking

In Chap. 2 it became clear that AIC values were relative rather than absolute for three reasons: (1) sample size impacted the size of AIC values, (2) there was the unknown constant $E_f[\log(f(x))]$ in the derivation of AIC from K–L information, and (3) some terms in the model set that are constant across models are often omitted. Simple differencing renders these issues moot. These differences, denoted as Δ_i, are standardized by the AICc value for the best model (the minimum AICc value). In fact, such differencing defines the best model as always having $\Delta_{best} \equiv 0$.

AICc Differences are Fundamental Units

Formally, the differences, Δ_i, are defined as

$$\boldsymbol{\Delta_i = AICc_i - AICc_{min}.}$$

These values are estimates of the expected K–L information (or distance) between the best (selected) model and the *i*th model. These differences apply when using AICc, QAICc (Sect. 6.2), or TIC, are on the scale of information, and are additive.

At this point we have science hypotheses and their associated models on a standard measurement scale. Although a scale of "information" might seem odd at first, it is little different than working with meters and kilometers or feet and miles.

Kullback–Leibler information is the distance from each of the models to full reality, whereas the Δ_i values relate to the distance between each of the models to the best one (Fig. 4.1). Everything is scaled to the best model where $\Delta_{best} \equiv 0$. This is convenient and is like so many other things in our experience. For example, in horse racing everything is scaled to the winner (quickest horse). The absolute time of the winning horse is unimportant because track conditions change from race to race and year to year. So, we speak of the winning horse and the second horse being two lengths behind, etc. Putting various science hypotheses in terms of the best one (i.e., the one with the most empirical support) is expected and should not be mistaken as arbitrary.

In a similar way, we measure the height of mountains and cities as the distance above sea level. Sea level has been convenient as a basis in scaling heights, much like Δ_{best} is convenient in scaling information and assessing the distance to other hypotheses and their models. Of course, such scaling to Δ_i values does not change the ranks based on AICc. It does make the examination of ranks visually easy as one merely looks for the model with $\Delta = 0$ and realizes that this is the model estimated to be the best (closest to truth). We must bear in mind that these are estimates and if we had a replicate data set of the same size and from the same process, a different hypothesis might be estimated to be the best in that case. We will quantify this uncertainty (called *model selection uncertainty*) using simple methods outlined in this chapter and Chap. 5.

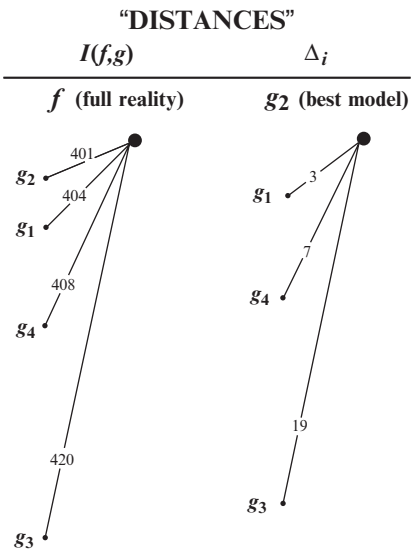

FIG. 4.1. Kullback–Leibler distances are with respect to the conceptual full reality (*left*), while the Δ_i are distances with respect to the best model and are on the scale of information.

An interesting issue arises because the Δ_i are on an information scale (Sect. 3.3.5). Under some fairly week assumptions it turns out that only science hypotheses and their models having Δ_i values in the range of 0 to perhaps 9–12 are plausible to most objective people (plausibility is a value judgment!). Even though AICc values for a particular problem might be in the 175,800–179,400 range, only models where Δ is within about 0 to 9 or 12, or 14 have much credibility. This will turn out to be a very useful result in practical application with real data. When I give the "window" above, I am deliberately trying to be vague about the upper bound as I do *not* want readers to consider some arbitrary cutoff (as in the α-level in testing theory). Science is about estimation and understanding; it is not about cutoffs or dichotomies.

Still, models with Δ values close to 0 have a lot of empirical support. Models with Δ values in the rough range 4–7 have considerably less support, whereas models with Δ values in the fringes (say 9–14) have relatively little support. Others, still further away, might be dismissed by most observers as *implausible*. The rationale for these rough guidelines is provided in Sect. 4.4; in addition, Royall (1997) offers similar guidelines. If observations are not independent but are assumed to be independent then these simple guidelines cannot be expected to hold. Likewise, if there are thousands of models, these guidelines may not hold entirely (however, if there are thousands or millions of models, then the endeavor is questionable anyway). The reader should not take these guidelines as inviolate as there are situations to which they may not apply well. Approaches to allow a more careful interpretation of the evidence are covered in the following material; thus, these rough guidelines are not necessary.

The Old Rule About $\Delta_i > 2$

One occasionally sees a rule that when a model has $\Delta_i > 2$ it is a poor model, or of little value, etc. This rule was seen in some early literature but I advise against it. Models with Δ_i values of 2 or 4 or even 6–7 or so have some meaningful support and should not be dismissed. In particular, the multimodel inference framework (Chap. 5) invites the use of information in such "second string" models.

The analysis is hardly finished once the ranking has been done and the best model has been estimated. One would examine and interpret the estimates of model parameters, the covariance matrix, and other aspects of the model estimated to be the best.

The absolute value of AICc is unimportant, it is the *differences* that can be directly related to *information*. These differences are an important quantity in the methods introduced immediately below.

4.2 Model Likelihoods

Parameter estimation under a likelihood framework is based on

$$\mathcal{L}(\underline{\theta}|\underline{x}, g_i),$$

meaning "the likelihood as a function of only the unknown parameters ($\underline{\theta}$), *given* the data ($\underline{x}$) and the particular model (g_i, such as binomial or normal or log-normal)." The likelihood is a function of the unknown parameters that must be estimated from the data (Appendix A). Still, the point is, this function allows the computation of likelihoods of various (tentative) parameter values. Likelihood values are relative and allow comparison. The objective is to find the parameter value that is most *likely* (i.e., the one that maximizes the likelihood) and use it as the MLE, the asymptotically best estimate of the unknown parameter, given the data and the model. With this background as a backdrop, I can introduce the concept of the likelihood of a model, given the data.

The Likelihood of Model *i*, Given the Data

The concept of the likelihood of the parameters, given the data and the model, i.e., $\mathcal{L}(\theta \mid \underline{x}, g_i)$ can be extended to the likelihood of model i given the data, hence $\mathcal{L}(g_i \mid \underline{x})$,

$$\mathcal{L}(g_i \mid \underline{x}) \propto \exp\left(-\tfrac{1}{2}\Delta_i\right).$$

Akaike suggested this simple transformation in the late 1970s. The likelihood of a model, given the data, offers the analyst a powerful metric in assessing the strength of evidence between *any* two competing hypotheses. This likelihood is very different from the usual one used in parameter estimation (obtaining MLEs). Both likelihoods are relative and useful in comparisons; they are not probabilities in any sense.

This simple expression allows one to compute the discrete likelihood of model i and compare that with the likelihood of other hypotheses and their models. Such quantitative evidence is central to empirical science. Chamberlin and Platt would greatly appreciate having the (relative) likelihood, based on the data, of each of his multiple working hypotheses. Notice that the $-\frac{1}{2}$ in the simple equation above merely removes the -2 that Akaike introduced in defining his AIC. Had he not used this multiplier, the likelihood of model i would have been just $\exp(\Delta_i)$ or $e^{\Delta i}$.

Likelihoods are relative and have a simple raffle ticket interpretation. One must think of likelihoods in a stochastic sense, such as the chance of winning a raffle based on the number of tickets the opponent has. Likelihoods are not like lifting weights, where the results are largely deterministic; if someone can lift considerably more than his/her opponent, the chances are good that he/she can do this again and again. There is little or no variation or "chance" in such activities. It is useful in evaluating science hypotheses to think in terms of the number of tickets each of the R hypotheses might have. If H_3 has a likelihood of 3 and H_5 has a likelihood of 300, then it is clear that evidence points fairly strongly toward support of hypothesis H_5 as it has 100 times the empirical support of hypothesis H_3 (we can say formally that it is 100 times more *likely*). Models having likelihoods of 3 and 300 are similar in principle to two people, one having three raffle tickets and the other having 300 tickets. Likelihoods are another way to quantify the strength of evidence between any model i and any other model j; there is no analogy with the "multiple testing problem" that arises awkwardly in traditional hypothesis testing. Computation of the model likelihoods is trivial once one has the Δ_i values.

4.3 Model Probabilities

Before proceeding to define model probabilities, we must define a relevant target value of such probabilities. Given a set of R models, representing R science hypotheses, one of these models is, in fact, the best model in the K–L information or distance sense. Like a parameter, we do not know which of the models in the set is actually *the* K–L best model for the particular sample size. Given the data, the parameters, the model set, and the sample size, one such model *is* the K–L best; we do not know which model is the best but we can estimate it using AICc. Moreover, we can estimate the uncertainty about our selection (our estimate of the model that is the best). This is crucial; we need a measure of the "model selection uncertainty." The target is not any notion of a "true model," rather the target is the actual best-fitted model in an expected K–L information sense. This concept must include the uncertainty in estimating the model parameters.

It is important not to confuse the "K–L best" model with a "good model." If all the models in the set are poor, these methods attempt to identify the best of these but in the end they all remain poor. Thus, one should examine such things as adj R^2, residual plots, and goodness-of-fit tests to be sure some of the models in the set are worthwhile.

It must be noted that the best model in a K–L sense depends on sample size. If n is small then the K–L best model will be of relatively low dimension. Conversely, if n is large, the K–L best model will be richer in structure and parameterization. These concepts tie back to the Principle of Parsimony (Sect. 2.3.4) and tapering effect sizes (Sect. 2.3.5). These are not easy concepts to grasp but are fundamental to model based inference. Oddly, the mathematical calculations here are trivial relative to the more difficult conceptual issues.

Model Probabilities

To better interpret the relative likelihoods of models, given the data and the set of R models, one can normalize these to be a set of positive "Akaike weights," w_i, adding to 1,

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R} \exp\left(-\frac{1}{2}\Delta_r\right)}.$$

These weights are also Bayesian posterior model probabilities (under the assumption of savvy model priors) and the formula applies when using AICc, QAICc (see Sect. 6.2), or TIC. A given w_i is the probability that model i is the expected K–L best model

$$w_i = \text{Prob}\left\{g_i \,\middle|\, \text{data}\right\}.$$

These probabilities are another weight of evidence in favor of model i as being the actual K–L best model in the candidate set. These w_i values are most commonly called model probabilities (given the model set and the data). These can be easily computed by hand, but a simple spreadsheet might avoid errors or too much rounding off. The term "Akaike weight" was coined because of their use in multimodel inference (Chap. 5) and before it was realized that these could be derived as Bayesian posterior model probabilities. The term *model probability* will often suffice.

The estimated K–L best model (let this best model be indexed by b) always has $\Delta_b \equiv 0$; hence, for that model $\exp(-(\frac{1}{2})\Delta_b) \equiv 1$. The odds for the ith model actually being the K–L best model is just $\exp(-(\frac{1}{2})\Delta_i)$. It is often convenient to reexpress such odds as the set of model probabilities as above.

The bigger a Δ_i is, the smaller the model probability w_i, and the less plausible is model i as being the actual K–L best model for full reality based on the sample size used. These Akaike weights or model probabilities give us a way to calibrate or interpret the Δ_i values; these weights also have other uses and interpretations (see below). While most forms of evidence are relative, model probabilities are absolute, conditional on the model set.

4.4 Evidence Ratios

In Sect. 4.1, some explanation was given to help interpret the Δ_i values. Evidence ratios can be used to make interpretation more rigorous and people should *not* use these as automatic cutoff values. Evidence is continuous and arbitrary cutoff points (e.g., $\Delta_i > 3$) should not be imposed or recognized.

Evidence Ratios

Evidence ratios between hypotheses i and j are defined as

$$E_{i,j} = \mathcal{L}(g_i \mid x) / \mathcal{L}(g_j \mid x) = w_i / w_j .$$

Evidence ratios are relative and not conditioned on other models in or out of the model set. Evidence ratios are trivial to compute; just a ratio of the model probabilities or model likelihoods.

Evidence ratios can be related back to the differences, Δ_i. An evidence ratio of special interest is between the estimated best model (min) and some other model i. Then the Δ value for that ith model can be related to the best model as

$$E_{\min,i} = w_{\min} / w_i = e^{(-(1/2)\Delta_i)}.$$

Evidence ratios also have a raffle ticket interpretation in qualifying the strength of evidence. For example, assume two models, A and B with the evidence ratio = $\mathcal{L}(g_A|\text{data})/\mathcal{L}(g_B|\text{data}) = 39$. I might judge this evidence to be at least moderate, if not strong, support of model A. This is analogous to model A having 39 tickets while model B has only a single ticket.

Some relevant values of Δ_i and the evidence ratio are given below.

Interpreting AICc Differences

The strength of evidence for the best model vs. any other model i is shown as a function of Δ_i:

Δ_i	Evidence ratio
2	2.7
4	**7.4**
6	20.1
8	**54.6**
10	148.4
11	244.7
12	403.4
13	665.1
14	1,096.6
15	1,808.0
16	**2,981.0**
18	8,103.1
20	22,026.0
50	72 billion

The second row, for example, is read as "if a model i has $\Delta_i = 4$, then the best model has 7.4 times the weight of evidence relative to model i (i.e., the best model has 7.4 raffle tickets whereas the other model has only one ticket). Using the information in the table above it is easy to see why models with Δ somewhere in the 8–14 range would be judged by most objective people as having little plausibility. Even odds of 55 to 1, (i.e., $\Delta = 8$) might often be judged as a "long shot." Models with delta >15–20 must surely be judged to be implausible.

A 7.4 to 1 advantage is pretty good and you might bet your used bike on a game with these odds; however, you must be careful as there is still a reasonably large chance (risk) that you would lose. Thus, a model where $\Delta = 4$ should not be dismissed; it has some reasonable empirical support. In contrast, a model with $\Delta = 16$ has but one ticket whereas the best model has almost 3,000 tickets (see table above). This is a clear case where you would not want to bet on the model where $\Delta = 16$; it is probably better to dismiss the model as being implausible! More extreme is the case where a model has $\Delta = 25$; here the odds of that model being, in fact, the best K–L model are remote (about 270,000 to 1) and most reasonable people might agree that the model should be dismissed. Still, an important point is that the evidence is the numerical value of the evidence ratio; this is where the objective science stops. Value judgments may follow and help interpret and qualify the science result. This is a good place to ask if the reader of this material is motivated to buy a ticket for a state or national lottery?

The table above makes it clear that models with Δ values below about 8 or 12 are in a window of some reasonable plausibility. Surely models with $\Delta >$ say 20 can probably be dismissed (unless the data are quite dependent (Sect. 6.2) or have been substantially compromised). No automatic cutoff is appropriate here; we must "qualify our mind to comprehend the meaning of evidence" as Leopold said in 1933. The "science answer" stops at the ranks, the model likelihoods, the model probabilities, and the evidence ratios. The *interpretation* involves a value judgment and can be made by anyone, including the investigator. Burnham and Anderson (2002:320–323) provide a more complicated example of these measures of evidence involving T_4 cell counts in human blood.

If the sample size is small or even of moderate size, care is needed in dismissing high-dimensional models as implausible. As sample size increases, additional effects can be identified. Often when sample size is small, there is a large amount of model selection uncertainty, keeping one from rejecting models with several parameters. This is another reason to design data collection such that sample size is as large as possible to meet objectives.

As Chamberlin pointed out, everyone wants a simple answer, even in cases where the best answer is not simple. Prior training in statistics has imprinted many of us with dichotomies that are in fact artificial and arbitrary (e.g., $P < 0.05$ in null hypothesis testing). However, in everyday life, people can live comfortably without such arbitrary cutoffs and rulings of "significance."

Consider a football score, 7 to 10. Most neutral observers would conclude the game was "close," but the team with 10 points won. No one bothers to ask if the win was "statistically significant." Of course, the winning team (hardly expected to be neutral) could claim (a value judgment) that they hammered their hapless opponents. However, alumni for the losing team might claim "last minute bad luck" or "poor refereeing" toppled their otherwise superior team. Still others might look at the number of yards rushing, the number of interceptions, and other statistics, and perhaps point to further interpretations of the evidence. In the end, perhaps all we can really say is that the game was close and if the teams played again under similar conditions, we could not easily predict the winner. This is a case where value judgments might vary widely; that is, the hard evidence is a bit thin to clearly suggest (an inference) which might be the better team.

Going further, two other teams play and the score is 3 to 35. Here, one must fairly conclude that the winning team was very dominating. The quantitative evidence is more clear in this case. Again, no issue about "statistical significance" is needed; the score (evidence) is sufficient in this case. The game was a "thumping" and any neutral observer could easily judge which team was better. If the two teams were to play again under similar conditions, one would suspect the winner could be successfully predicted (an inference). In this case, value judgments would probably vary little from person to person, based on the evidence (the score).

Summarizing, in football, the evidence is the final score and in science, evidences are things like model probabilities and evidence ratios. Interpretation of this evidence involves value judgment and these might (legitimately) vary substantially or little at all. Scientists should avoid arbitrary dichotomies and "cutoffs" in interpreting the quantitative evidence.

4.5 Hardening of Portland Cement

Here we return to the example of the hardening of Portland cement from Sects. 2.2.1 and 3.7 to illustrate the nature of scientific evidence

Model	K	$\hat{\sigma}^2$	log $\mathcal{L}$	AICc	Δ_i	w_i
{mean}	2	208.91	−34.72	71.51	39.1	0.0000
{12}	4	4.45	−9.704	32.41	0.0	0.9364
{12 1*2}	5	4.40	−9.626	37.82	5.4	0.0629
{34}	4	13.53	−16.927	46.85	14.4	0.0007
{34 3*4}	5	12.42	−16.376	51.32	18.9	0.0000

Readers should verify the simple computations in the last two columns of the table above. For example, using $\mathbf{\Delta_i = AICc_i - AICc_{min}}$, $\Delta_1 = 71.51 - \mathbf{32.41} = 39.1$, $\Delta_2 = 32.41 - \mathbf{32.41} = 0$ and $\Delta_3 = 37.82 - \mathbf{32.41} = 5.4$. The computation of the model probabilities (w_i) is made from

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta_r\right)}$$

and this is more simple than it might appear. The first step is to tabulate the values of $\exp(-(\frac{1}{2})\Delta_i)$ for $i = 1, 2, \ldots, 5$ as these are needed for both numerator and denominator

$$\begin{aligned}
&\exp\left(-\tfrac{1}{2}\Delta_1\right) = \exp\left(-\tfrac{1}{2}\cdot 39.1\right) = 0.0000,\\
&\exp\left(-\tfrac{1}{2}\Delta_2\right) = \exp\left(-\tfrac{1}{2}\cdot 0\right) = 1,\\
&\exp\left(-\tfrac{1}{2}\Delta_3\right) = \exp\left(-\tfrac{1}{2}\cdot 5.4\right) = 0.0672,\\
&\exp\left(-\tfrac{1}{2}\Delta_4\right) = \exp\left(-\tfrac{1}{2}\cdot 14.4\right) = 0.0007,\\
&\exp\left(-\tfrac{1}{2}\Delta_5\right) = \exp\left(-\tfrac{1}{2}\cdot 18.9\right) = 0.0001.
\end{aligned}$$

More decimal places should often be carried; I will give the results to four places beyond the decimal (thus, 0.0000 is not meant to reflect exactly zero, it is only zero to four places). The quantity $\sum_{r=1}^{R}\exp(-(1/2)\Delta_r)$ in the denominator is merely the sum of these five numbers, **1.068**. Finally, the model probabilities are w_1 = 0.0000/**1.068** = 0.0000, w_2 = 1/**1.068** = 0.936, w_3 = 0.0672/**1.068** = 0.063, and so on.

Until this chapter, we could only rank the five hypotheses and their models; this, of course, allowed the estimated best model (i.e., model {12}) to be identified. Now, we have the ability to see that two of the models can be judged to be implausible: models {mean} and {34 3∗4} (even the better of these two (i.e., model {34 3∗4}) has a model probability of only 0.00008) and might be dismissed as the set evolves. Only models {12} and {12 1∗2} have noticeable empirical support whereas model {34} has very little empirical support (model probability of 0.0007). The model probabilities are exactly what Chamberlin would have wanted, but it must be remembered that they are conditional on the model set.

What is the evidence for the interaction 1∗2? An evidence ratio answers this question: E = 0.9364/0.0629 = 14.9 (from the table just above). This measure indicates that the support for the model without the interaction is nearly 15 times that of the model with the interaction. What would be your value judgment, based on the evidence, in this case? In other words, how would you qualify the result concerning the interaction term?

Additional evidence here is to look at the MLE for the beta parameter (β_3) for the interaction term. We find, $\hat{\beta}_3$ = 0.0042 with an approximate 95% confidence interval of (−0.020, 0.033). One must conclude that there is little support for the 1∗2 interaction term. Notice also that the deviance (deviance

= −2 × log $\mathcal{L}$) changed little as the interaction term was added: 19.408 vs. 19.252, again making support for the interaction term dubious (this is the "pretending variable" problem, Sect. 3.6.8).

Let us imagine a member of the cement hardening team had always favored model {34} and felt strongly that it was superior to the rest. What support does the member have, based on this small sample of 13 observations? The evidence ratio of the best model vs. model {34} is 1,339 to 1: not much support of model {34}. He must quickly try to argue that the data were flawed, measurements were in error, etc. No reasonable observer will overturn odds of over 1,300:1; the evidence is strongly against the member's belief and all reasonable value judgments would confirm this.

4.6 Bovine Tuberculosis in Ferrets

The addition of the information-theoretic differences and model probabilities allow more evidence to be examined:

Hypotheses	K	log $\mathcal{L}$	AICc	Rank	Δ_i	w_i
H_1	6	–70.44	154.4	4	50.8	0.0000
H_2	6	–986.86	1,987.2	5	1,883.6	0.0000
H_3	6	–64.27	142.1	3	38.5	0.0000
H_4	6	–45.02	103.6	1	0.0	0.7595
H_5	6	–46.20	105.9	2	2.3	0.2405

Here, it seems clear that the evidence strongly suggests that hypotheses 1–3 are implausible; the probability that H_3 is, in fact, the K–L best model is less than 4×10^{-8} and the other two models have far less probability. This finding certainly allows the set to evolve to the next level. Support of hypotheses 4 and 5 is somewhat tied, with H_4 having the edge by a factor of about three times (i.e., 0.7595/0.2405 ≈ 3) the support over H_5. One cannot rule out the support for H_5 based on the evidence (Fig. 4.2).

Note that these model probabilities are conditional on the set of five hypotheses and the five model probabilities sum to 1. One can compute evidence ratios among any of the five, even if one or two of the hypotheses are deleted. Often the interest is in evidence ratios with the best model vs. some other model; however, one is free to select any models i and j for evaluation by simple evidence ratios. Note that because all models here have the same number of estimable parameters, the penalty term can be ignored in this particular case and one can use just the deviance (−2 × log $\mathcal{L}$) as "AICc." Finally, note that if, for some reason, model g_2 is dropped from the set, then the other five model results must be renormalized to sum to 1 (in this particular example it would make no difference).

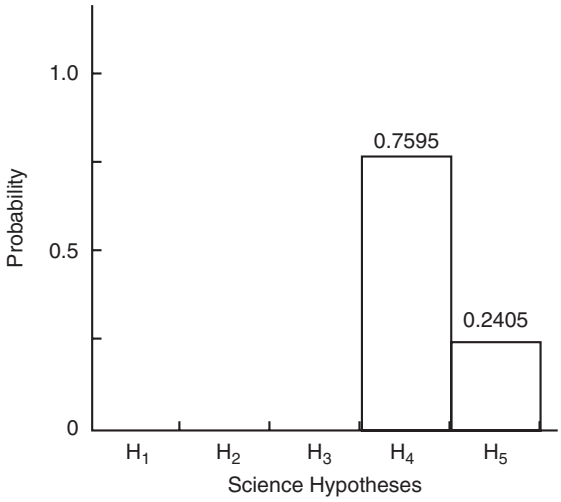

FIG. 4.2. Model probabilities for the five hypothesis concerning transmission of bovine tuberculosis in feral ferrets in New Zealand.

4.7 Return to Flather's Models and R^2

Anderson and Burnham (2002:94–96) presented a reanalysis of nine models for the species accumulation curves for data from Indiana and Ohio from Flather (1992, 1996). The science objective here was to explore the structure of the accumulation process and allow predictions concerning species accumulation and richness. Hypotheses concerning species accumulation were modeled as nonlinear regressions, and one model was found to be quite better than the other eight models. Here we will look at an interesting side issue as this gives some insights into the power of these newer methods.

The adjusted R^2 value in regression analysis measures the proportion of the variance in the response variable that is in common with variation in the predictor variables: it is a measure of the overall "worth" of the model, at least in terms of within-sample prediction. Adj R^2 for the nine models ranged from 0.624 to 0.999 (Table 4.1).

Examining Table 4.1 shows that five of the models had adj R^2 values above 0.980. One might infer that any of these five models must be quite good; even the worst model with $R^2 = 0.624$ might appear pretty good to many of us. Going further, two of the models (models 8 and 9) had an adj $R^2 > 0.999$; surely these models are virtually tied for the best and are both excellent models for a structural description of the accumulation process and for prediction. Although the statements above certainly seem reasonable, they are misleading. For example, the evidence ratio for the (estimated) best model (model 9) vs. the second best model (model 8) is

$$E_{\min,8} = w_{\min} / w_8 = w_9 / w_8 = \exp\left((\Delta_8 / 2) = \exp(163.4 / 2) \approx 3.0 \times 10^{35}\right)$$

TABLE 4.1. Summary of nine *a priori* models of avian species-accumulation curves from Flather (1992, 1996). The models are shown in order according to the number of parameters (K); however, this is only for convenience.

Model	K	log $\mathcal{L}$	AICc	Δ_i	w_i	adj R^2
1. ax^b	3	−110.82	227.64	813.12	0.0000	0.962
2. $a + b \log x$	3	−42.78	91.56	677.04	0.0000	0.986
3. $a(x/(b + x))$	3	−172.20	350.40	935.88	0.0000	0.903
4. $a(1 - e^{-bx})$	3	−261.58	529.17	1114.65	0.0000	0.624
5. $a - bc^x$	4	−107.76	223.53	809.01	0.0000	0.960
6. $(a + bx)/(1 + cx)$	4	−24.76	57.53	643.01	0.0000	0.989
7. $a(1 - e^{-bx})^c$	4	25.42	−42.85	542.63	0.0000	0.995
8. $a(1 - [1 + (x/c)^d]^{-b})$	5	216.04	−422.08	163.40	0.0000	0.999
9. $a[1 - e^{-(b(x-c))^d}]$	5	297.74	−585.48	0	1.0000	0.999

K is the number of parameters in the regression model plus 1 for σ^2
Some values of the maximized log-likelihood are positive because some terms were constant across models and were omitted.
The first eight model probabilities are 0 to at least 34 decimal places.

and is very convincing evidence that even the second best model has no plausibility (note, too, $\Delta_8 = 163.4$). The evidence ratio for model 4 (where adj $R^2 = 0.624$) is 1.1×10^{242}.

Adjusted R^2 values are useful as a measure of the proportion of the variation "in common" but are not useful in model selection (McQuarrie and Tsai 1998). Information-theoretic approaches show convincingly that all of the models are incredibly poor, relative to model 9 (the best model). This is a case where there is essentially no model selection uncertainty; the actual K–L best model among those considered, beyond any doubt, is model 9. Of course, everything is conditional on the set of hypotheses and their models. It is possible that a tenth model might be better yet; this is why the science objective is to let these sets evolve as more is learned. On the other hand, if model 9 was never considered, then model 8 would look very good relative to the other models. There are several reasons why adjusted R^2 is poor in model selection; its usefulness should be restricted to description.

4.8 The Effect of a Flood on European Dippers

Marzolin's data on the European dipper (*Cinclus cinclus*) have become a classic for illustrating various analytical issues (see Lebreton et al. 1992). Rationale for the hypotheses was outlined in Sect. 2.7 and this would be a point where this introductory material should be reread.

The problem focused on two models, one model without a flood effect on survival (model $\{\phi,p\}$) and another with a flood effect ($\{\phi_n,\phi_f,p\}$), where ϕ is

the apparent annual probability of survival, p is the annual probability of capture, and the subscripts denote a normal year (n) or a flood year (f). Under an information-theoretic approach, the results can be given very clearly in terms of the model probabilities (the w_i):

Model	K	AICc	Δ_i	Probability, w_i
$\{\phi,p\}$	2	670.866	4.706	0.0868
$\{\phi_n,\phi_f,p\}$	3	666.160	0.000	0.9132

The table above provides the quantitative evidence and the evidence favors the hypothesis that annual survival decreased in the two years when a flood occurred. The evidence ratio is 10.5 = 0.9132/0.0868 indicating that the flood-affect hypothesis had about ten times more empirical support than the hypothesis that survival did not change during the years of the flood. [*Note*: I am careful to avoid any notion of causation here; however, the result can be said to be confirmatory because of the *a priori* hypotheses.]

Looking deeper, the MLEs for ϕ and measures of precision for the two models are:

Model		MLE	95% confidence interval
$\{\phi,p\}$		0.5602	0.5105–0.6888
$\{\phi_n,\phi_f,p\}$	(n)	0.6071	0.5451–0.6658
	(f)	0.4688	0.3858–0.5537

The MLEs of the capture probability for the models were 0.9026 vs. 0.8998, respectively; these are virtually identical. The estimated "effect size" for survival was 0.1383 (=0.6071 − 0.4688), s.e. = 0.0532, and 95% confidence interval of (0.0340, 0.2425). The "evidence" in this case includes the model probabilities, an evidence ratio, estimates of effect size, and measures of precision. These entities allow a qualification of the hard, quantitative evidence.

In Sect. 2.7, we thought deeper about the issues and considered the following, alternative hypotheses:

- Was there a survival effect just the first year of the flood $\{\phi_{f1},\phi_{nf},p\}$?
- Or just the second year of the flood $\{\phi_{f2},\phi_{nf},p\}$?
- Or was the recapture probability (p) also effected by the flood $\{\phi_f,\phi_{nf},p_f,p_{nf}\}$?
- Or even $\{\phi,p_f,p_{nf}\}$ where survival was not impacted, but the recapture probabilities were?

We could now address these questions as if they were *a priori*; they should have been as they are questions begging to be asked. Instead, for example, we will assume the candidate set was not well thought out and, these four new hypotheses arose *post hoc* (a not unusual situation). Thus, we will admit the

TABLE 4.2. Model selection results for the European dipper data.

Model	K	log $\mathcal{L}$	AICc	Δi	wi	Rank	"Tickets"
$\{\phi_n,\phi_f,p\}$	3	**−330.051**	666.160	0.000	**0.559**	1	1000[a]
$\{\phi_f,\phi_{nf},p_f,p_{nf}\}$	4	**−330.030**	668.156	1.996	**0.206**	2	369
$\{\phi_{f1},\phi_{nf},p\}$	3	−331.839	669.735	3.575	0.094	3	167
$\{\phi_{f2},\phi_{nf},p\}$	3	−332.141	670.338	4.178	0.069	4	124
$\{\phi,p\}$	2	−333.419	670.866	4.706	0.053	5	95
$\{\phi,p_f,p_{nf}\}$	3	−333.412	672.881	6.721	0.019	6	35

[a]This column is just 1,000 × exp(-(1/2)Δ_i) to illustrate the evidence for each of the six hypotheses in terms of how many raffle tickets each had. This is just a useful way to comprehend the evidence. Ticket numbers help in understanding; I am not proposing these be placed in publications

post hoc nature of the issue and examine a set of six models (the two original *a priori* models and the four new *post hoc* models). We will be prepared to tell our reader what we did and promise to treat these *post hoc* results as more tentative and speculative. The results are summarized in rank order in Table 4.2.

Note that the evidence ratio between model $\{\phi,p\}$ and model $\{\phi_n,\phi_f,p\}$ did not change (0.559/0.053 = exp(4.706/2) = 10.5) as the four new models were added; however, the model probabilities (w_i) related to the candidate set change as models are added or deleted. Model probabilities are conditional on the candidate set.

The key question deals with a possible flood affect on the capture probabilities (p_i). The evidence ratio for $\{\phi_n,\phi_f,p\}$ vs. $\{\phi_n,\phi_f,p_f,p_{nf}\}$ = 0.559/0.206 = 2.7. This evidence might be judged as weak; nonetheless, it does not support the notion that the capture probabilities varied with the year of the flood. Another piece of evidence lies in the MLEs and their profile likelihood intervals (Appendix A) for model $\{\phi_n,\phi_f,p_f,p_{nf}\}$:

Parameter	Estimate	95% profile interval
p_{nf}	0.904	(0.836, 0.952)
p_f	0.893	(0.683, 0.992)

Overall, most would judge the evidence to be largely lacking; survival seems to have been influenced by the flood, but not the capture probabilities. It may also be interesting to note that the two models with no flood effect on survival were ranked last. If a much larger data set had been available, perhaps other effects would have been uncovered (i.e., the concept of tapering effect size, Sect. 2.3.5).

More careful examination of the table suggests another example of the "pretending variable" problem with model $\{\phi_f,\phi_{nf},p_f,p_{nf}\}$. The addition of one additional parameter did not improve the fit as the log-likelihood value changed very little (−330.051 vs. −330.030) and the model was penalized by only Δ = 2 units). This is further evidence that the flood had little effect on capture probabilities.

The data on dippers were taken by gender, thus science hypotheses could have related gender to the flood. For example, are females more likely to have lowered survival in flood years as perhaps they are tending the nest close to the water's edge? Perhaps the capture probability is higher for males as they forage over a wider area and more vulnerable to being caught. These and other questions are the type of thinking and hypothesizing that Chamberlin would have wanted – especially if it were done *a priori* to data analysis.

Although the dipper data and the science questions are fairly simple, they illustrate several key points. First, we wish to make an *inductive inference* from the sample data to the population of dippers along the streams in the years sampled. Second, that inference is *model based*. Third, the model probabilities and evidence ratios admit a high degree of rigor in the inferences. This approach refines Platt's (1964) concept of strong inference.

The data here are a series of 0s and 1s indicating the capture history of each dipper; simple plots of these histories would reveal nothing of interest. That is, the data are capture histories for each bird:

11001101
10001101
00100001
00000010

for four birds in this example (see Appendix B). This is a clear case where inference must be model based as simple graphics would not be informative. Fourth, the models are products of multinomial distributions, and MLEs and their covariance matrix are derived from statistical theory. Fifth, the initial investigation was *confirmatory* and this allowed a more directed analysis and result. I chose to assume that the four additional hypotheses were *post hoc* although, ideally, they too would have been the result of *a priori* thinking and hypothesizing. Finally, perhaps no hypotheses (above) would warrant dismissal; no hypothesis seems to be virtually without support. Compare this situation with that from the disease transmission in ferrets where 2–3 of the hypotheses could easily be dismissed. Thus the model set for dippers might evolve by refinement of the existing models (not likely) or by further hypothesizing, but not by dropping some hypotheses and their models. Data on this population have been collected for several additional years and those data could be subjected to the same six hypotheses and analysis methods; at that time the candidate set might begin to evolve more rapidly.

4.9 More About Evidence and Inference

I find it useful to think about evidence about a parameter as the maximum likelihood estimate $\hat{\theta}$ and its profile likelihood interval (a type of confidence interval, see Appendix A). Both of these evidential quantities are dependent

on a model. Traditionally, the model was assumed to be *given*. Now we have rigorous methods to select the model from the data. In Chap. 5, we will see that estimates of model parameters can be made from all the models in the set (multimodel inference) and this often has distinct advantages.

Less well known, but equally important, are the types of evidences about alternative science hypotheses.

Types of Evidence

There are three main kinds of evidences, in addition to simple ranking, concerning the evaluation of alternative science hypotheses:

1. Model probabilities, the probability that model i is, in fact, the K–L best model. These are denoted as w_i and they are also formal Bayesian posterior model probabilities.
2. The (relative) likelihood of model i, given the data. Such likelihoods are denoted as $\mathcal{L}(g_i|\text{data})$. Likelihoods are always relative to something else of interest; e.g., a likelihood of 0.31 means nothing by itself.
3. Evidence ratios provide the empirical evidence (or support) of hypothesis i vs. j, $E_{ij} = w_i / w_j$; simply the ratio of model likelihoods or model probabilities for any two models i and j.

The evidence ratio relates to any models i and j, regardless of other models in or out of the set. Model probabilities depend on exactly R hypotheses in the set; they are conditional on the set being fixed. Of course, all the three quantities stem from the differences, Δ_i. It is the Δ_i that are on a scale of information and are the basis for the other measures of evidence.

A simple ranking of alternative science hypotheses is a form of evidence – ranking is based on the data and stems from "the information lost when using a hypothesis or model to approximate full reality." We want models that keep information loss to a minimum, as seen in Chap. 3.

None of these types of evidences are meant to be used in a dichotomous yes/no fashion. These are ways to quantify the evidence; this is where the science stops. From there, one can qualify the evidence to aid in understanding and interpretation. One should avoid thinking that, for example, "Δ_4 is greater than 10, therefore it is unimportant or implausible" (or worse yet, "not significant"). There are always value judgments in interpreting evidence; often, virtually every objective person might arrive at the same value judgment, whereas in other cases, considered opinion (judgment) will vary substantially.

An important component of the analysis of data remains model assessment of the global model. Here, R^2, goodness-of-fit evaluations, and analysis of residuals have a role in assuring that some of the models are useful. When this is in doubt, it is sometimes useful to include a model with little or no structure and be sure that it is relatively implausible compared to, say, the global model.

4.10 Summary

Going back to 1890, Chamberlin asked, "What is the measure of probability on one side or the other?" Although it took nearly 100 years to develop methods to provide such probabilities, they are very useful in the evaluation of alternative science hypotheses. When data are entered into a proper likelihood and $-\log \mathcal{L}$ is computed, the units are "information," regardless of the original measurement units associated with the data.

I suspect Chamberlin might have been fairly content (in 1890) with a way to merely rank alternative hypotheses, based on the empirical data. Ranking is so interesting and compelling. Still, I find the concept of evidence ratios to be very central in evaluating the "strength of evidence" for alternative science hypotheses. This preference is not intended to downplay the model probabilities and Bayesians would certainly favor these in their framework. To me, the evidence ratios are so revealing and so comparative; however, I use the other quantities also.

It is conceptually important to recognize that the Δ_i define a narrow window of values likely to be judged as plausible. This window might reasonably be defined as 0 to 8–13 or so. This window exists regardless of the scale of measurement (e.g., millimeters or kilometers), the type of response variable (e.g., continuous, binomial), dimensionality of the parameter vector (3–8 or maybe 20–100), nested or nonnested models, and number of variables at hand. [This window assumes independence of the data (outcomes) – but see Sect. 6.2. Time series and spatial models provide a proper treatment for these dependent data.]

Every new method described in this chapter is simple to compute and understand; they also seem compelling. It is very important that scientists understand their results at a deep level. Biologists working in a team situation may often find that others on the team "did the analysis" and they have little or no idea as to what was done and for what reason. These are methods that should be relatively easy to comprehend and this is central to good science. Good science strategy tries to push the information gained to let the set evolve. This ever-changing set is the key to rapid advancement in knowledge and understanding.

Perhaps the biggest drawback to these approaches, as with all approaches, is the challenge to carefully define and model the hypotheses. If a model poorly reflects the science hypothesis, then everything is compromised. This is a continual challenge; I think statistics courses for both majors and nonmajors could better focus on modeling, rather than the current emphasis on null hypothesis testing.

Still the focus, from a science standpoint, must be on the alternative hypotheses. This focus is so central but so easily distracted. Investigators running "all possible models" via some powerful computer software have missed the entire point of the study. Models should arise to represent carefully derived science hypotheses; they should not arise just because the software makes them pos-

sible. Finally, it should be noted that the information-theoretic approach unifies parameter estimation and the selection of a parsimonious approximating model; both can be viewed as optimization problems.

4.11 Remarks

The idea of the likelihood of the model, given the data, was suggested many years ago by Akaike (e.g., Akaike 1978b, 1979, 1980, 1981b; also see Bozdogan 1987; Kishino 1991).

Royall's (1997) book focused on the Law of Likelihood and likelihood ratios as evidence for one hypothesis over another. His treatment was useful in cases for simple models and where the number of parameters in the two models is the same, or perhaps differ by 1. This short book offers many valuable insights and is easy to follow, but is not a book for beginners.

It is helpful to recall that we are not just trying to fit the data; instead, we are trying to recover the information in the data and allowing robust prediction by using a model. It is model based inference and that inference comes from a *fitted* model (i.e., the parameters have been estimated from the data, given that model). These realities are important while trying to avoid the notion that a model is true and fully represents reality.

AICc formulates the problem explicitly as a problem of *approximation* of reality. Real data do not come from models. We cannot hope to find full reality using models and finite data. As a crutch, we can think of full reality as infinite dimensional; however, full reality is unlikely to be parameterized. Parameters are a construct useful in many science contexts, but many parts of full reality are not even parameterized. A "good" model successfully separates information from "noise" or noninformation in the data, but never fully represents truth.

If a person insists that they have credible prior beliefs about each model being the actual K–L best model, then these beliefs can be formally specified as a set of discrete prior probabilities (in a Bayesian sense) on models, ζ_i. It seems unlikely that we would have a prior belief in this case as this would entail a belief about approximations to full reality as well as the expected parsimonious trade-offs in model fitting and how this varies by sample size. The ζ_i must be the prior probabilities as to which model, *when fit to the data* (θ is estimated), is best for representing the (finite) information in the data. If one had a set of priors on models (ζ_i), then there is a simple way to generalize the model probabilities to reflect this:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)\zeta_i}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta_r\right)\zeta_r}.$$

This is not a Bayesian approach and its properties are unknown (maybe unknowable?). Neither Burnham nor I would use this approach, but it does exist if people can honestly think they have informed beliefs about the model priors ζ_i. More detail on this approach is given by Burnham and Anderson (2002:76–77).

It is clear now why Akaike considered the information-theoretic methods to be extensions of Fisher's likelihood theory. People believing in and using these methods might be classed as "likelihoodists" to distinguish themselves from traditional "frequentists" with their emphasis on null hypothesis testing and notions of confidence intervals and "Bayesians" with their priors on parameters and priors on models and inference being based on the posterior distribution.

Classifications such as this are sometimes helpful but perhaps the best philosophy is to use the best tool for a specific science problem. I do not hide the fact that I think the best, and most practical, approach is most often the objective one based on K–L information and entropy. I do not support the subjective aspects of the Bayesian methods in science but believe the objective Bayesian approach is well suited for the large class of random effects models (some additional comments are offered in Appendix D).

One often hears of "traditional" frequentist statistics – meaning null hypothesis testing approaches and frequentist confidence intervals, etc. It should be noted that although such approaches date back to the early parts of the twentieth century, the Bayesian approach goes back quite further, to the mid-1700s.

I do not favor the word "testing" regarding these new approaches, as it seems confusing and might imply testing null hypotheses. "Evaluate" seems a better choice of words. The word "significant" is best avoided in all respects.

Fisher tried to avoid a cutoff value to interpret his *P*-values. He once chose $\alpha = 0.05$ saying that a one in 20 chance might be something to suggest interest. Neyman and Pearson's approach called for α to be set in advance and was used in decisions about "significance" and power and so on. People have a natural tendency to simplify, but the reliance on (always arbitrary) cutoff points is to be discouraged in the empirical sciences.

The so-called Akaike weights are the basis for approaches to making inference from all the models in the set (Chap. 5), but they are interestingly more. Ken Burnham found that if one takes the Bayesian approach that Schwarz (1978) took and uses "savvy priors" on models instead of vague priors on models, AIC drops out (instead of BIC)! There is more to this at a technical level, but it is in this sense that it can be shown that AIC can be derived from Bayesian roots (Burnham and Anderson 2004). Thus, the w_i are properly termed Bayesian posterior model probabilities. I usually prefer to call them just "model probabilities."

Bayesians are struggling with the issue of assigning prior probabilities on models and how to make these "innocent" or vague. It seems reasonable to investigate further the general notion of savvy or K–L priors on models in a Bayesian framework (see Burnham and Anderson 2004 for a more technical treatment of this issue).

The "theory of the theory" can get very deep in model selection (e.g., Linhart and Zucchini 1986; van der Linde 2004, and especially Massart 2007). Also

see Vol. 50 of the *Journal of Mathematical Psychology*, edited by Wagenmakers and Waldorp, for some current research results.

4.12 Exercises

1. Reread the information in Sect. 2.3.1 on beak lengths in Darwin's finches. There were seven science hypotheses concerning possible evolutionary changes in beak lengths and these were represented by seven models. Your research team has collected data on beak lengths over many years and it is time for an analysis. Your technician has studied each of the models and has obtained the MLEs of the model parameters and the value of the maximized log-likelihood function. These are summarized below:

Model i	log £	K	AICc	Δ_i	w_i
1	−66.21	2			
2	−57.77	5			
3	−59.43	6			
4	−60.98	6			
5	−49.47	6			
6	−49.47	7			
7	−49.46	8			

   You are asked to complete the computations for the remaining columns and use the results as evidence in addressing the following questions:

   a. What hypothesis is best supported? Why do you say it is the "best"?
   b. Do these data provide evidence for two phenotypes? Why do you say this?
   c. Is the evidence for two phenotypes strong? Weak? Be specific.
   d. What is the supporting evidence for the covariates? Which one? Both? Why?
   e. What is the evidence for the interaction term in model 7?
   f. What further, *post hoc*, analyses might be considered? How would future research be molded by these results?

2. Some results concerning the affect of a flood on dippers were given in Sect. 4.8. The two models were nested and a likelihood ratio test was made: model $\{\phi,p\}$ vs. model $\{\phi_n,\phi_f p\}$. The test statistic was 6.735 on 1 degree of freedom, giving a P-value of 0.0095. This is markedly different (by an order of magnitude!) from the probability of model $\{\phi,p\}$, given the data (0.0868). Explain this issue in some detail.
3. The nightly weather forecast uses the words probability of rain and likelihood of rain as synonyms. Detail the technical differences in these terms.
4. You are given the result, "the probability of hypothesis H_4, represented by its model g_4 is 0.53." This result might be called "conditional," but how? Explain. How is this different from an evidence ratio?

5. Can you intuit why adjusted R^2 leads to overfitted models? (advanced question)
6. Can you prove to yourself that the differencing leading to the Δ_i values removes the constant (across models) term that was omitted in the heuristic derivation from K–L information to AIC?
7. You have a colleague in Spain collaborating with you on an interesting science problem. You have worked together and have settled on six hypotheses, have derived six models to carefully reflect these hypotheses, and have fitted the data to the models. Thus, you have the MLEs, the covariance matrix, and other associated quantities. Everything seems to be in order. She has taken the lead in the analysis and provides the following AICc values:

H_1 3,211	H_4 14,712
H_2 3,230	H_5 7,202
H_3 3,234	H_6 5,699

   She tentatively suggests that H_1 is the best, H_2 and H_3 are very close, but that H_4–H_6 are very poor (implausible, actually). Compose your e-mail comments on her findings so far. Form your response in terms of strength of evidence and ways to provide this.
8. While the mapping from the residual sum of squares (RSS) to the maximized log(L) is simple, many people stumble in trying to use the information-theoretic approaches in a simple "*t*-test" or ANOVA setting. First, consider a paired design of sample size n. (a) Lay out the models for the null and alternative hypotheses and display the computations for the RSS. (b) Lay out the same items for the unpaired design. Then, discuss how one would proceed to compute AICc, Δ_i, Prob$\{H_0|data\}$, Prob$\{H_A|data\}$ and an evidence ratio. Finally, compare the advantages and differences between the usual *t*-statistics, the *P*-value, and rulings of statistical "significance" vs. model probabilities and evidence ratios.