CHAPTER 24

MODEL COMPARISON

"Entities are not to be multiplied without necessity" - - - William of Occam, *ca* 1340

We have seen in some detail how to conduct inferences – test hypotheses, estimate parameters, predict future observations – within the context of a preassigned model, representing some working hypothesis about the phenomenon being observed. But a scientist must be concerned also with a bigger problem; how to decide between different models when both seem able to account for the facts.

Stated very broadly, the problem is hardly new; some 650 years ago the Franciscan Monk William of Occam perceived the logical error in the Mind Projection Fallacy.† This led him to teach that some religious issues might be settled by reason, but others only by faith. He removed the latter from his discourse, and concentrated on the areas where reason might be applied – just as Bayesians seek to do today when we discard mind–projecting mythology (such as assertions of limiting frequencies in experiments that have never been performed), and concentrate on the things that are meaningful in the real world. His famous epigram quoted above, generally called "Occam's razor", represents a good start on the principles of reasoning that he needed, and that we still need today. But it was also so subtle that only through modern Bayesian analysis has it been well understood.

Of course, from our present vantage point it is clear that this is really the same problem as that of compound hypothesis testing, considered already in Chapter 4. Here we need only generalize that treatment and work out further details, but some extra care is needed. As long as we work within a single model, normalization constants tend to cancel out and so need not be introduced at all. But when two different models appear in a single equation, the normalization constants do not cancel out, and it is imperative that all probabilities be correctly normalized.

Formulation of the Problem

To see why this happens, recall first what Bayes' theorem tells us about parameter estimation. A model M contains various parameters denoted collectively by θ. Given data D and prior information I, and assuming the correctness of model M, to estimate the parameters we first apply Bayes' theorem:

$$p(\theta|D, M, I) = p(\theta|M, I)\,\frac{p(D|\theta, M, I)}{p(D|M, I)} \qquad (24\text{–}1)$$

in which the denominator serves as the normalizing constant:

$$p(D|M, I) = \int p(D, \theta|M, I)\, d\theta = \int p(D|\theta, M, I)\, p(\theta|M, I)\, d\theta \qquad (24\text{–}2)$$

† Occam's position, stated in the language of his time, was that "Reality exists solely in individual things, and universals are merely abstract signs." Translated into Twentieth Century language: the abstract creations of the mind are not realities in the external world. Unfortunately for him, some of the cherished 'realities' of contemporary orthodox theology were just the things to which he denied reality; so this got him into trouble with the Establishment. But later religious scholars who disregarded Occam's teachings and tried doggedly to apply reason in problems not amenable to it, produced only nonsense. Evidently, Occam was a forerunner of modern Bayesians, to whom all this sounds very familiar.

which we see is the prior expectation of the likelihood $L(\theta) = p(D|\theta, M, I)$; that is, its expectation over the prior probability distribution $p(\theta|M, I)$ for the parameters.

Now we move up to a higher level problem; to judge, in the light of the prior information and data, which of a given set of different models $\{M_1 \cdots M_r\}$ is most likely to be the correct one. Bayes' theorem gives the posterior probability for the j'th model as

$$p(M_j|D, I) = p(M_j|I)\,\frac{p(D|M_j, I)}{p(D|I)}\,, \qquad\qquad 1 \le j \le r\,. \tag{24–3}$$

But we may eliminate the denominator $p(D|I)$ by calculating instead odds ratios as we did in Chap. 4. The posterior odds ratio for model M_j over M_k is

$$\frac{p(M_j|D, I)}{p(M_k|D, I)} = \frac{p(M_j|I)}{p(M_k|I)} \cdot \frac{p(D|M_j, I)}{p(D|M_k, I)} \tag{24–4}$$

and we see that the same probability $p(D|M_j, I)$ that appears in the single–model parameter estimation problem (24–1) only as a normalizing constant, now appears as the fundamental quantity determining the status of model M_j relative to any other.[‡] The exact measure of what the data have to tell us about this, is always the prior expectation of its likelihood function, over the prior probability $p(\theta_j|M_j, I)$ for whatever parameters θ_j may be in that model (they are generally different for different models). All probabilities must be correctly normalized here, otherwise we are violating our basic rules and the likelihood ratio in (24–4) is arbitrary nonsense even when it is not zero or infinite.

Intuitively, the model favored by the data is the one that assigns the highest probability to the observed data, and therefore "explains the data" best. This is just a repetition, at a higher level, of the likelihood principle for parameter estimation within a model.

But it is not yet clear how an Occam principle can emerge from this. Indeed, the principle has never been stated in exact, well–defined terms. Later writers have tried, almost universally, to interpret it as saying that the criterion of choice is the 'simplicity' of the competing models, although it is not clear that Occam himself used that term. Centuries of discussion by philosophers trying to make this interpretation brought no appreciable clarification of what is meant by 'simplicity'.[⋆] We think that concentration of attention exclusively on that undefined term has prevented understanding of the real point, which is merely that a model with unspecified parameters is a composite hypothesis, not a simple one. For this reason some interesting new features appear, arising from the internal structure of the parameter space.

Parameters Known in Advance: To see this, suppose first that there is no such internal space; the parameters of a model are known exactly ($\theta = \theta'$) in advance. This amounts to assigning a prior $p(\theta_j|M_j I) = \delta(\theta_j - \theta'_j)$, whereupon (24–2) reduces to

[‡] This logical structure is more general even than the Bayesian formalism; we shall see in Volume 2 that it persists in the pure maximum–entropy formalism, where in statistical mechanics the relative probability P_j/P_k of two different phases, such as liquid and solid, is the ratio of their partition functions Z_j/Z_k, which are the normalization constants for the sub–problems of prediction within one phase, although they are not expectations of any likelihoods. In Bayesian analysis, the data are indifferent between two models when their normalization constants become equal; in statistical mechanics the temperature of a phase transition is the one at which the two partition functions become equal. In chemical thermodynamics it is customary to state this as equality of the "free energies" $F_j \simeq \log Z_j$. This illustrates the basic unity of Bayesian and maximum–entropy reasoning, in spite of their superficial differences.

[⋆] Indeed, for a time the notion of 'simplicity' was given up for dead, because of the seeming impossibility of defining it. The tedious details are recounted by Rosenkrantz (1977).

$$p(D|M_j, I) = p(D|\theta'_j, M_j, I) = L_j(\theta'_j) \tag{24–5}$$

just the likelihood of θ'_j within the j'th model. Evidently, this will be a maximum if θ'_j happens to be equal to the maximum likelihood estimate $\hat{\theta}_j$ for that model and the data. Then the posterior odds ratio (24–4) would reduce to

$$\frac{p(M_j|D,I)}{p(M_k|D,I)} = \frac{p(M_j|I)}{p(M_k|I)} \cdot \frac{(L_j)_{max}}{(L_k)_{max}} \tag{24–6}$$

This is the conventional Bayes' theorem result of Chapter 4. If the parameters were known to have, for each model, the most favorable values for the given data set, each model becomes in effect a simple hypothesis rather than a composite one [this is almost self–contradictory, for if the data were different one would have to suppose also different prior information about the parameters in order to retain (24–6)].

But this extreme case is also very unrealistic; usually, the parameters are unknown and in the problems 'amenable to reason' where useful inferences are possible, the data D will be more informative about the parameters within some model M_j than is the prior information; that is, as a function of θ_j the likelihood function $L_j(\theta_j) = p(D|\theta_j, M_j, I)$ will be more sharply peaked, at some point picked out by the data, than is the prior probability $p(\theta_j|M_j, I)$. Then in the exact integral (24–2) most of the contribution will come from a "high likelihood region" Ω comprising a small neighborhood of that sharp peak. There is hardly any loss of generality in assuming this because, unless it is true, we would consider the data too meager to permit any useful new inferences and although the Bayesian procedure would still be valid in principle we would, like Occam, 'remove the problem from our discourse' as being unproductive.

Parameters Unknown: Consider for the moment only the j'th model and drop the index j. Let there be m parameters $\theta \equiv \{\theta_1 \cdots \theta_m\}$ in the model and expand $\log L(\theta)$ about the maximum likelihood point $\hat{\theta} = \{\hat{\theta}_1 \cdots \hat{\theta}_m\}$:

$$\log L(\theta_1 \cdots \theta_m) = \log L_{max} + \frac{1}{2} \sum_{i,j=1}^{m} \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) + \cdots \tag{24–7}$$

Then near the peak a good approximation is a multivariate gaussian function:

$$L \simeq L_{max} \exp\left[-\frac{1}{2}(\theta - \hat{\theta})' \,\Sigma^{-1} (\theta - \hat{\theta})\right] , \tag{24–8}$$

with the "inverse covariance matrix"

$$(\Sigma^{-1})_{i,j} \equiv - \left[\frac{\partial^2 \log L}{\partial \theta_i \, \partial \theta_j}\right]_{\theta = \hat{\theta}} \tag{24–9}$$

Our supposition is that the prior density does not vary appreciably over the high likelihood region Ω; then if we were estimating the parameters θ_j, in the approximation (24–8) we should be led to estimates of the form

$$(\theta_j)_{est} = \hat{\theta}_j \pm \sqrt{\Sigma_{jj}} \tag{24–sam}$$

and the integral (24–2) is

$$p(D|M,I) \simeq L_{max}\,(2\pi)^{m/2}\,\sqrt{\det(\Sigma)}\,p(\hat{\theta}|M,I) \tag{24–10}$$

Let us interpret this result in terms of parameter space volumes. We may define the high likelihood region Ω more explicitly by the conditions that:

(1) It is as compact as possible; within Ω the likelihood everywhere exceeds some nominal threshold value L_0. The volume of this region is then

$$V(\Omega) = \int_{L>L_0} d\theta_1 \cdots d\theta_m \ . \tag{24–11}$$

(2) The integrated likelihood should be given by $L_{max}\,V(\Omega)$:

$$L_{max}\,V(\Omega) = \int L(\theta)\,d^m\theta = L_{max}\,(2\pi)^{m/2}\,\sqrt{\det\Sigma}\ . \tag{24–12}$$

Then a rectangular function equal to L_{max} on Ω, zero elsewhere, is a crude approximation to the likelihood function and it has, in the present approximation, the same implications for model comparison as does the actual likelihood function. These conditions determine the effective high–likelihood volume of parameter space without any need to calculate the threshold L_0:

$$V(\Omega) = (2\pi)^{m/2}\,\sqrt{\det\Sigma} \tag{24–13}$$

Note that this is just the normalization constant for the above multivariate gaussian function;†

$$\int \exp\left[-\frac{1}{2}\,(\theta-\hat{\theta})'\,\Sigma^{-1}\,(\theta-\hat{\theta})\right]\cdot\frac{d\theta_1\cdots d\theta_m}{V(\Omega)} = 1\ . \tag{24–14}$$

Exercise (24.1). Evaluate the threshold L_0 and the dimensions $(\theta-\hat{\theta})$ of the high–likelihood region Ω by direct evaluation of the integral (24–11). Note that the matrix Σ, being real, symmetric, and positive definite, can be diagonalized: $\Sigma = U\,\Lambda\,U^{-1}$, where U is an $(m\times m)$ real orthogonal matrix; that is, its transpose is $U' = U^{-1}$ so $\det(U) = \pm 1$, and $\Lambda_{ij} = \lambda_i\,\delta_{ij}$ is the diagonalized matrix. Now make the "spherical" change of variables from $\{\theta_1\cdots\theta_m\}$ to $\{x_1\cdots x_m\}$, where

$$(\theta_k - \hat{\theta}_k) = \sum_{i=1}^{m} U_{ki}\sqrt{\lambda_i}\,x_i$$

and perform the integrations in the x–space. Show that, in x–space, the region Ω is the interior of an m–dimensional sphere of radius $R \simeq (m/e)^{1/2}$; and that the exact volume of such a sphere is $\pi^{m/2}\,R^m/(m/2)!$. As a check, in the cases $m = (1, 2, 3)$ this reduces to $(2R,\ \pi R^2,\ 4\pi R^3/3)$ respectively, as it should.

The right–hand side of (24–10) now becomes $L_{max}\,V(\Omega)\,p(\hat{\theta}|M,I)$. But since by hypothesis $p(\theta|M,I)$ does not vary appreciably over Ω, the quantity

† Indeed, the maximum density for any continuous distribution $p(x_1\cdots x_n)$ is dimensionally the reciprocal of an n-dimensional volume, which can always be interpreted as the volume of a high–probability region.

$$W \equiv \frac{1}{L_{max}} \int L(\theta) p(\theta|MI)\, d^m\theta \simeq V(\Omega)\, p(\hat{\theta}|M, I) \qquad (24\text{–}15)$$

is for all practical purposes just the *amount of prior probability* contained in the high likelihood region Ω, and our fundamental model comparison rule now becomes

$$\frac{p(M_j|D,I)}{p(M_k|D,I)} = \frac{p(M_j|I)}{p(M_k|I)} \cdot \frac{(L_j)_{max}}{(L_k)_{max}} \cdot \frac{W_j}{W_k} \qquad (24\text{–}16)$$

in which we see revealed, by comparison with (24–6), the Occam factor (W_j/W_k) arising from the internal parameter spaces of the models. In (24–16), the likelihood factor depends only on the data and the model, while the Occam factor depends also on the prior information about its internal parameters. If two different models achieve the same likelihoods $(L_j)_{max}$, then in sampling theory terms they account for the data equally well, and one would think that we have no basis for choice between them. Yet Bayes' theorem tells us that there is an another quality in the models; the prior information which may still give strong grounds for preference of one over the other. Indeed, the Occam factor may be so strong that it reverses the likelihood judgment.

But Where is the Idea of Simplicity?

The relation (24–16) has much meaning that unaided intuition could not (or at least, did not) see. If the data are highly informative compared to the prior information, then the relative merit of two models is determined by the product of two factors;

(1) How high a likelihood can be attained on the parameter space of a model?

(2) How much prior probability is concentrated in the high–likelihood region Ω picked out by the data?

But neither of these seems concerned with the simplicity of the model (which seems for most of us to refer to the number of different assumptions that are made – for example, the number of different parameters that are introduced – in defining a model).

To understand this, let us ask: "How do we all decide these things intuitively?" Having observed some facts, what is the real criterion that leads us to prefer one explanation of them over another? Suppose that two explanations, A and B, could account for some proven historical facts equally well. But A makes four assumptions, each of which seemed to us already highly plausible; while B makes only two assumptions, but they seem strained, far–fetched, and unlikely to be true. Every historian finds himself in situations like this; and he does not hesitate to opt for explanation A, although B is intuitively simpler. Thus our intuition asks, fundamentally, not how *simple* the hypotheses are; but rather how *plausible* they are.

But there is a loose connection between simplicity and plausibility, because the more complicated a set of possible hypotheses, the larger the manifold of conceivable alternatives, and so the smaller must be the prior probability of any particular hypothesis in the set.

Now we see why 'simplicity' could never be given a satisfactory definition (that is, a definition that accounted in a satisfactory way for these inferences); it was a poorly chosen word, directing one's attention away from the essential component of the inference. For generations, scientists and philosophers have opined vaguely that 'simple hypotheses are more plausible' without giving any logical reason why this should be so. We suggest that this should be turned around: we should say rather that 'more plausible hypotheses tend to be simpler'. An hypothesis that we consider simpler is one that has fewer equally plausible alternatives.

None of this could be comprehended at all within the confines of orthodox statistical theory, whose ideology did not allow the concept of a probability for a model or for a fixed but unknown parameter. Orthodoxy tried to compare models entirely in terms of their different sampling distributions, which took no note of *either* the simplicity of the model or the prior information! But it was unable to do even this, because then all the parameters within a model became 'nuisance parameters' and that same ideology denied one any way to deal with them.‡ Thus orthodox statistics was a total failure on this problem, and this held up progress for most of the 20'th Century.

It is remarkable that, although the point at issue is trivial mathematically, generations of mathematically competent people failed to see it because of that conceptual mindset which kept thinking doggedly in terms of simplicity. But once the point is seen, it seems intuitively obvious and one cannot comprehend how 'simplicity' could ever have been imagined as the criterion. This just reminds us again that the human brain is an imperfect reasoning device; although it is fairly good at drawing reasonable conclusions, it often fails to give a convincing rationale for those conclusions. For this we really do need the help of probability theory as logic.

Thus, although Bayes' theorem always gives us the correct answer to whatever question we ask of it, it often does this in such a slick, efficient way that we are left bewildered and not quite understanding how it happened. The present problem is a good example of this, so let us try to understand the situation better intuitively. Denote by M_n a model for which $\theta = \{\theta_1, \ldots, \theta_n\}$ is n–dimensional, ranging over a parameter space S_n. Now introduce a new model M_{n+1} by adding a new parameter θ_{n+1} and going to a new parameter space S_{n+1}, in such a way that $\theta_{n+1} = 0$ represents the old model M_n. We shall presently give an explicit calculation with this scenario, but first let us think about it in general terms.

On the subspace S_n the likelihood is unchanged by this change of model; $p(D|\theta, M_{n+1}, I) = p(D|\theta, M_n, I), \quad \theta \in S_n$. But the prior probability $p(\theta|M_{n+1}, I)$ must now be spread over a larger parameter space than before and will, in general, assign a lower probability to a neighborhood Ω of a point in S_n than did the old model.

For a reasonably informative experiment, we expect that the likelihood will be rather strongly concentrated in small subregions $\Omega_n \in S_n$ and $\Omega_{n+1} \in S_{n+1}$. Therefore, if with M_{n+1} the maximum likelihood point occurs at or near $\theta_{n+1} = 0$, Ω_{n+1} will be assigned less prior probability than is Ω_n with model M_n, and we have $p(D|M_n, I) > p(D|M_{n+1}, I)$; the likelihood ratio generated by the data will favor M_n over M_{n+1}. This is the Occam phenomenon.

Thus, if the old model is already flexible enough to account well for the data, then as a general rule Bayes' theorem will, like Occam, tell us to prefer the old model. It is intuitively simpler if by 'simpler' we mean a model that occupies a smaller volume of parameter space, and thus *restricts us to a smaller range of possible sampling distributions.* Generally, the inequality will go the other way only if the maximum likelihood point is far from $\theta_{n+1} = 0$ (*i.e.* a significance test would indicate a need for the new parameter), because then the likelihood will be so much smaller on Ω_n than on Ω_{n+1} that it more than compensates for the lower prior probability of the latter; as noted, Occam would not disagree.

But intuition does not tell us at all, quantitatively, how great this discrepancy in likelihoods must be in order to bring us to the point of indifference between the models. Furthermore, having seen this mechanism, it is easy to invent cases (for example, if the introduction of the new parameter is accompanied by a redistribution of prior probability on the old subspace S_n) in which Bayes' theorem may contradict Occam because it is taking account further circumstances undreamt of in Occam's philosophy. So we need specific calculations to make these things quantitative.

An Example: Linear Response Models

Now we give a simple analysis that illustrates the above conclusions and allows us to calculate definite numerical values for the likelihood and Occam factors. We have a data set

‡ This and other criticisms of orthodox hypothesis testing theory were made long ago by Pratt (1961).

$D \equiv \{(x_1, y_1) \cdots (x_n, y_n)\}$ consisting of measured values of (x, y) in n pairs of observations. We may think of x as the 'cause' and y as the 'effect' although this is not required. For the general relations below the 'independent variables' x_i need not be uniformly spaced or even monotonic increasing in the index i. From these data and any prior information we have, we are to decide between two conceivable models for the process generating the data. For model M_1 the responses are, but for irregular measurement errors e_i, linear in the cause:

$$M_1: \qquad y_i = \alpha x_i + e_i\,, \qquad 1 \le i \le n \tag{24–100}$$

while for model M_2 there is also a quadratic term:

$$M_2: \qquad y_i = \alpha x_i + \beta x_i^2 + e_i\,, \tag{24–101}$$

which represents, if β is negative, an incipient saturation or stabilizing effect (if β is positive, an incipient instability). We may think, for concreteness, of x_i as the dose of some medicine given to the i'th patient, y_i as the resulting increase in blood pressure. Then we are trying to decide whether the response to this medicine is linear or quadratic in the dosage. But this mathematical model applies equally well to many different scenarios.† Whichever model is correct, the errors of measurement of y_i are supposed to be the same, and we assign a joint sampling distribution to them:

$$p(e_1 \cdots e_n|I) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{e_i^2}{2\sigma^2}\right\} = \left(\frac{w}{2\pi}\right)^{n/2} \exp\left\{-\frac{w}{2}\,\Sigma_i e_i^2\right\} \tag{24–102}$$

where $w \equiv 1/\sigma^2$ is the 'weight' parameter, more convenient in calculations than σ^2.

Digression: The Old Sermon Still Another Time: Again, we belabor the meaning of this, as discussed in Chapter 7. In orthodox statistics, a sampling distribution is always referred to as if it represented an 'objectively real' fact, the frequency distribution of the errors. But we doubt whether anybody has ever seen a real problem in which one had prior knowledge of any such frequency distribution, or indeed prior knowledge that any limiting frequency distribution exists; how could one ever acquire information about the long–run results of an experiment that has not been performed? That is part of the Mind–Projecting Mythology that we discard.

We recognize, then, that assigning this sampling distribution is only a means of describing our own *prior state of knowledge* about the measurement errors. The parameter σ indicates the general magnitude of the errors that we expect; The prior information I might, for example, be the variability observed in past examples of such data; or in a physics experiment it might not be the result of any observations, but rather obtained from the principles of statistical mechanics, indicating the level of Nyquist noise for the known temperature of the apparatus.

---

† For example, x_i might be the amount of ozone in the air in the i'th year, y_i the average temperature in January of that year. Or, x_i may be the amount of some food additive ingested by the i'th Canadian rat, y_i the binary decision whether that rat did or did not develop cancer. Or, x_i may be the amount of acid rain falling on Northern Germany in the i'th year, y_i the number of pine trees that died in that year; and so on. In other words, we are now in the realm of what were called 'linear response models' in the Preface, and the results of these calculations have a direct bearing on many currently controversial health and environmental issues. Of course, many real problems will require more sophisticated models than we are considering now; but having seen this simple calculation it will be clear how to generalize it in many different ways.

In particular, the absence of correlations in (24–102) is not an assertion that no correlations exist in the real data; it is only a recognition that we have no knowledge of such correlations, and therefore to suppose correlations of either sign is as likely to hurt as to help the quality of our inferences. Thus in one sense, by being noncommittal about it, we are only being honest and frankly acknowledging our ignorance. But in another sense, we are taking the safest, most conservative course; using a sampling distribution which will yield reasonable results whether or not correlations actually exist. But if we knew of any such correlations, we would be able to make still better inferences (although not much better) by use of a sampling distribution which contains them.

The reason for this is that correlations in a sampling distribution tell the robot that some regions of the vector sample space are more likely than others even though they have the same error magnitudes; then some details of the data that it would have to dismiss as probably noise, can be recognized as providing further evidence about systematic effects in the model.

Back to the Problem: The sampling distribution for model M_1 is then

$$M_1 : \qquad p(D|\alpha M_1) = \left(\frac{w}{2\pi}\right)^{n/2} \exp\left\{-\frac{w}{2}\sum_{i=1}^{n}(y_i - \alpha x_i)^2\right\} \tag{24–103}$$

The sum is

$$\Sigma = \sum_i (y_i^2 - 2\alpha y_i x_i + \alpha^2 x_i^2) = n(\overline{y^2} - 2\alpha\overline{xy} + \alpha^2\overline{x^2}) \tag{24–104}$$

where the bars denote, as before, averages over the data. The maximum likelihood estimate (MLE) of α is then found from $\partial\Sigma/\partial\alpha = n(-2\overline{xy} + 2\alpha\overline{x^2}) = 0$, or,

$$\alpha = \hat{\alpha} \equiv \frac{\overline{xy}}{\overline{x^2}} \tag{24–105}$$

which in this case is also called the 'ordinary least squares' (OLS) estimate. The likelihood (24–103) for model M_1 is then

$$L(\alpha, M_1) = \left(\frac{w}{2\pi}\right)^{n/2} \exp\left\{-\frac{nw}{2}\left[\overline{y^2} - \hat{\alpha}^2\overline{x^2} + \overline{x^2}(\alpha - \hat{\alpha})^2\right]\right\} \tag{24–106}$$

and we note in passing that, if we were using this to estimate α from the data, our result would be

$$(\alpha)_{est} = \hat{\alpha} \pm \frac{1}{\sqrt{nw\overline{x^2}}} \tag{24–joe}$$

Now, using (24–105), the 'global' sampling distribution for model M_1 in (24–3) contains two factors:

$$p(D|M_1 I) = \int p(D|\alpha M_1)\, p(\alpha|M_1 I)\, d\alpha = L_{max}(M_1)\cdot W \tag{24–107}$$

where

$$L_{max}(M_1) = \left(\frac{w}{2\pi}\right)^{n/2} \exp\left\{-\frac{nw}{2}\left(\overline{y^2} - \hat{\alpha}^2\overline{x^2}\right)\right\} = \left(\frac{w}{2\pi}\right)^{n/2} \exp\left\{\frac{nw}{2}\,\frac{\overline{xy}^2 - \overline{x^2}\,\overline{y^2}}{\overline{x^2}}\right\} \tag{24–108}$$

$$W_1 = \int \exp\left\{-\frac{nw}{2}\,\overline{x^2}(\alpha - \hat{\alpha})^2\right\} p(\alpha|M_1 I)\, d\alpha\,. \tag{24–109}$$

Now we are obliged to use a normalized prior for α; almost always it will be known that $|\alpha|$ cannot be enormously large (else there would be such a catastrophe that we would not be concerned with this problem); but we would seldom have any more specific prior information about it. We can indicate this by assigning a prior density

$$p(\alpha|M_1 I) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left\{-\frac{\alpha^2}{2\delta^2}\right\} \tag{24–110}$$

which says that we do not know whether α is positive or negative, but it is highly unlikely that $|\alpha|$ is much greater than δ. As we saw in Chapter 6, when we are estimating parameters within a single given model and have such vague prior information about them, the exact analytical form of the prior makes no difference in the conclusions. That is, the effect of different reasonable priors first appears in our conclusions in perhaps the tenth decimal place; but since we are calculating those final conclusions only to three or four decimal places, the effect of different priors is not just negligibly small; it is strictly nil. All priors that are essentially equal to a constant $C = p(\hat{\alpha}|M, I)$ over the region Ω of high likelihood, lead to the same conclusions; even the value of the constant C cancels out. But when we are comparing different models, C does not cancel out; it expresses the prior range (in this case, $2\delta \simeq 1/C$) of values that α might have. Then we are free to choose a Gaussian analytical form which makes it easy to do the integrations. Indeed, this choice can be justified also as representing the actual state of knowledge that we have in real problems. Then the likely error σ of our measurements is so much smaller than δ that over the high likelihood region Ω, the prior density for α is essentially constant and equal to $(2\delta)^{-1}$. Had we chosen a rectangular prior with width 2δ, it would have led to just the same result.

With the prior (24–110) we can do the integration (24–109) exactly, with the result

$$W_1 = \frac{1}{\sqrt{1 + nw\overline{x^2}\delta^2}} \exp\left\{-\frac{nw\overline{x^2}\,\hat{\alpha}^2}{2(1 + nw\overline{x^2}\delta^2)}\right\} \tag{24–111}$$

But this can be simplified greatly. In the first place, we see from (24–joe) that the accuracy with which the experiment can measure α is $\sigma/\sqrt{n\overline{x^2}}$, and δ is surely at least 100 times this, so

$$nw\overline{x^2}\delta^2 = n\overline{x^2}\frac{\delta^2}{\sigma^2} = \left(\frac{\text{prior range for } \alpha}{\text{accuracy of the measurement of } \alpha}\right)^2 \tag{24–112}$$

and this is typically very large numerically, of the order of 10^4 or greater. Therefore (24–111) may be written

$$W_1 = \frac{1}{\sqrt{nw\overline{x^2}\delta^2}} \exp\left\{-\frac{\hat{\alpha}^2}{2\delta^2}\right\} \tag{24–113}$$

But now δ is surely also at least 100 times greater than $\hat{\alpha}$; so $\hat{\alpha}^1/2\delta^2$ is less than 10^{-4}; and the Occam factor reduces, to all the accuracy we could use, simply to

$$W_1 = \frac{\text{accuracy of } \alpha \text{ measurement}}{\text{prior range for } \alpha} \tag{24–114}$$

***************** MORE COMING HERE! ***********************