“…the very game…”

A Tutorial on Mathematical Modeling

Michael P. McLaughlin
www.geocities.com/~mikemclaughlin

PREFACE

"OK, I've got some data. Now what?"

It is the quintessence of science, engineering, and numerous other disciplines to make quantitative observations, record them, and then try to make some sense out of the resulting dataset. Quite often, the latter is an easy task, due either to practiced familiarity with the domain or to the fact that the goals of the exercise are undemanding. However, when working at the frontiers of knowledge, this is not the case. Here, one encounters unknown territory, with maps that are sometimes poorly defined and always incomplete.

The question posed above is nontrivial; the path from observation to understanding is, in general, long and arduous. There are techniques to facilitate the journey but these are seldom taught to those who need them most. My own observations, over the past twenty years, have disclosed that, if a functional relationship is nonlinear, or a probability distribution something other than Gaussian, Exponential, or Uniform, then analysts (those who are not statisticians) are usually unable to cope. As a result, approximations are made and reports delivered containing conclusions that are inaccurate and/or misleading.

With scientific papers, there are always peers who are ready and willing to second-guess any published analysis. Unfortunately, there are as well many less mature disciplines which lack the checks and balances that science has developed over the centuries and which frequently address areas of public concern. These concerns lead, inevitably, to public decisions and warrant the best that mathematics and statistics have to offer, indeed, the best that analysts can provide. Since Nature is seldom linear or Gaussian, such analyses often fail to live up to expectations.

The present tutorial is intended to provide an introduction to the correct analysis of data. It addresses, in an elementary way, those ideas that are important to the effort of distinguishing information from error. This distinction, unhappily not always acknowledged, constitutes the central theme of the material described herein.

Both deterministic modeling (univariate regression) as well as the (stochastic) modeling of random variables are considered, with emphasis on the latter since it usually gets short shrift in standard textbooks. No attempt is made to cover every topic of relevance. Instead, attention is focussed on elucidating and illustrating core concepts as they apply to empirical data. I am a scientist, not a statistician, and these are my priorities.

This tutorial is taken from the documentation included with the Macintosh software package *Regress+* which is copyrighted freeware, downloadable at

http://www.geocities.com/~mikemclaughlin/software/Regress_plus.html

Michael P. McLaughlin
McLean, VA
October, 1999

For deeds do die, however nobly done,
And thoughts of men do as themselves decay,
But wise words taught in numbers for to run,
Recorded by the Muses, live for ay.

—E. Spenser [1591]

“…the very game…”

HIS mother was a witch (or so they said). Had she not poisoned the glazier’s wife with a magic drink and cursed her neighbors’ livestock so that they became mad? Did she not wander about the town pestering people with her recipes for medicines? Yet, in spite of such proofs, his filial devotion did not go unrewarded and, with the aid of a good lawyer plus the support of his friend and patron Rudolph II, Emperor of the Romans, King of Germany, Hungary, Bohemia, &c., Archduke of Austria, &c., the old woman made her final exit with less flamboyance than some of His Holy Imperial Majesty’s subjects might have wished. However, this is not about the mother but the son—and Mars. Johannes Kepler is remembered, to this day, for his insight and his vision. Even more than his contemporary, Galileo, he is honored not just for what he saw but because he invented a new way of looking.

In astronomy, as in most disciplines, how you look determines what you see and, here, Kepler had a novel approach. He began with data whereas all of his predecessors had begun with circles. The Aristotelian/Ptolemaic syllogism decreed that perfect motion was circular. Heavenly bodies were perfect. Therefore, they moved in circles, however many it took to save appearances.

It took a lot. When, more than a quarter of a century before Kepler, Nicolaus Copernicus finally laid down his compass, he had, quite rightly, placed the Sun in the center of the remaining seven known bodies but he had also increased the number of celestial circles to a record forty-eight!

Kepler commenced his intellectual journey along the same path. Indeed, in those days, it was the only path. After many false starts, however, he realized that a collection of circles just would not work. It was the wrong model; the data demanded something else. Kepler bowed to Nature and, without apology, substituted ellipses for circles. He was the first scientist to subjugate theory to observation in a way that we would recognize and applaud.

Of course, Kepler was in a unique position. Thanks to Tycho Brahe, he had the best data in the world and he was duly impressed. Still, he could have voted the party line and added yet more circles. Sooner or later, he would have accumulated enough parameters to satisfy every significant figure of every measurement. But it was wrong and it was the wrongness of it that impressed Kepler most of all. Although he drew his inspiration from the ancient Pythagoreans and the religious fervor of his own time, his words leave little doubt of his sincerity:

> “…Now, as God the maker played,
> he taught the game to Nature
> whom he created in his image:
> taught her the very game
> which he played himself…”

Kepler searched through the data and found the game. He learned the rules and showed that, if you played well enough, sometimes even emperors take notice. Today, our motivation is different but the game goes on. We begin, of course, with data.

DATA

There are two kinds of data: measurements and opinions. This discussion will focus exclusively on the former. In fact, although there are useful exceptions in many disciplines, here we shall discuss only quantitative measurements. Adopting this mild constraint provides two enormous advantages. The first is the advantage of being able to speak very precisely, yielding minimal concessions to the vagaries of language. The second is the opportunity to utilize the power of mathematics and, especially, of statistics.

Statistics, albeit a discipline in its own right, is primarily an ever-improving cumulation of mathematical tools for extracting information from data. It is information, not data, that leads ultimately to understanding. Whenever you make measurements, perform experiments, or simply observe the Universe in action, you are collecting data. However, real data always leave something to be desired. There is an open interval of quality stretching from worthless to perfect and, somewhere in between, will be your numbers, your data. Information, on the other hand, is not permitted the luxury of imperfection. It is necessarily correct, by definition. Data are dirty; information is golden.

To examine data, therefore, is to sift the silt of a riverbed in search of gold. Of course, there might not be any gold but, if there is, it will take some knowledge and considerable skill to find it and separate it from everything else. Not only must you know what gold looks like, but you also have to know what sorts of things masquerade as gold. Whatever the task, you will need to know the properties of what you seek and what you wish to avoid, the chemistry of gold and not-gold. It is through these properties one can be separated from the other.

An example of real data is shown in Table 1 and Figure 1.[1] This dataset consists of values for the duration of daytime (sunrise to sunset) at Boston, Massachusetts over three years. The first day of each month has been tabulated along with the longest and shortest days occurring during this period. Daytime has been rounded off to the nearest minute.

What can be said about data such as these? It can be safely assumed that they are correct to the precision indicated. In fact, Kepler's data, for analogous measurements, were much more precise. It is also clear that, at this location, the length of the day varies quite a bit during the year. This is not what one would observe near the Equator but Boston is a long way from tropical climes. Figure 1 discloses that daytime is almost perfectly repetitive from year to year, being long in the (Northern hemisphere) Summer and short in the Winter.

Such qualitative remarks, however, are scarcely sufficient. Any dataset as rich as this one deserves to be further quantified in some way and, moreover, will have to be if the goal is to gain some sort of genuine understanding. With scientific data, proof of understanding implies the capability to make accurate predictions. Qualitative conclusions are, therefore, inadequate.

Quantitative understanding starts with a set of well-defined *metrics*. There are several such metrics that may be used to summarize/characterize any set of N numbers, y_i, such as these daytime values. The most common is the *total-sum-of-squares*, TSS, defined in Equation 1.

---

[1] see file *Examples:Daytime.in* [FAM95]

Table 1. Daytime—Boston, Massachusetts (1995-1997)

Daytime (min.)	Day	Date
545	1	1 Jan 1995
595	32	
669	60	
758	91	
839	121	
901	152	
915	172	21 Jun 1995
912	182	
867	213	
784	244	
700	274	
616	305	
555	335	
540	356	22 Dec 1995
544	366	1 Jan 1996
595	397	
671	426	
760	457	
840	487	
902	518	
915	538	21 Jun 1996
912	548	
865	579	
782	610	
698	640	
614	671	
554	701	
540	721	21 Dec 1996
545	732	1 Jan 1997
597	763	
671	791	
760	822	
839	852	
902	883	
915	903	21 Jun 1997
912	913	
865	944	
783	975	
699	1005	
615	1036	
554	1066	
540	1086	21 Dec 1997
545	1097	1 Jan 1998

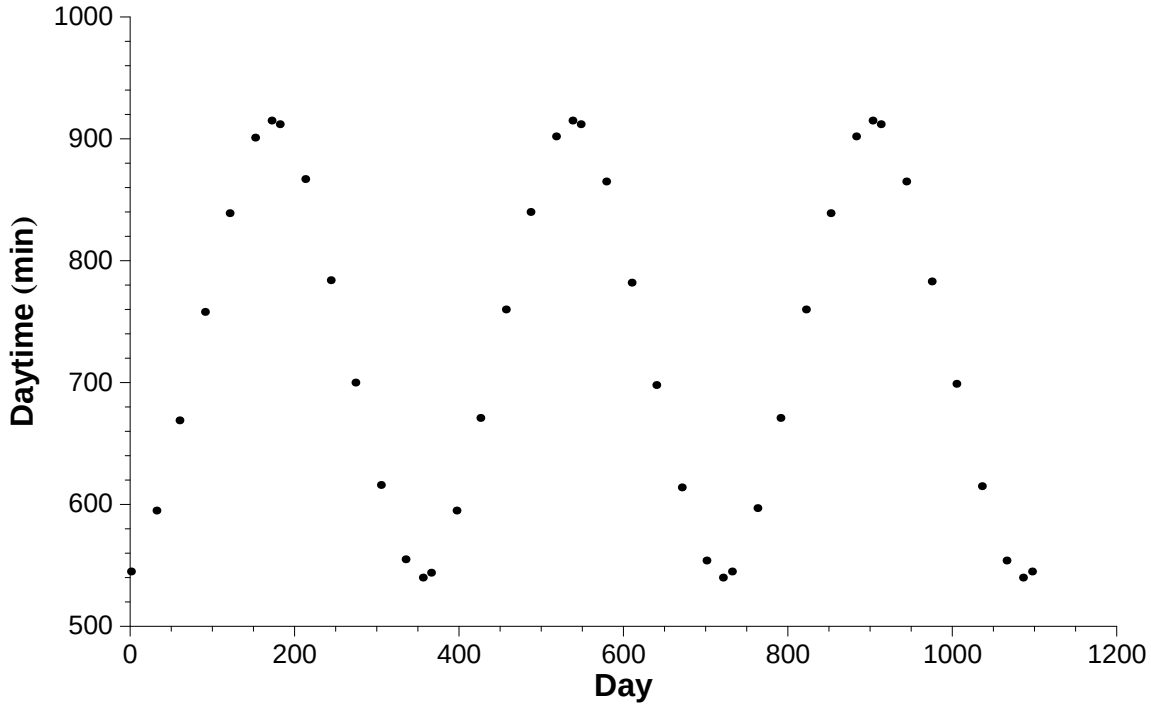

Figure 1. Raw Daytime Data

$$\text{Total-sum-of-squares} \equiv \text{TSS} = \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2 \qquad 1.$$

where $\bar{y}$ is the average value of y.

TSS is a positive number summarizing how much the y-values vary about their average (*mean*). The fact that each x (day) is paired with a unique y (daytime) is completely ignored. By discounting this important relationship, even a very large dataset may be characterized by a single number, i.e., by a *statistic*. The average amount of TSS attributable to each point (Equation 2) is known as the *variance* of the variable, y. Lastly, the square-root of the variance is the *standard deviation*, another important statistic.

$$\text{Variance of y} \equiv \text{Var}\left(y\right) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2 \qquad 2.$$

In Figure 1, the y-values come from a continuum but the x-values do not. More often, x is a continuous variable, sampled at points chosen by the observer. For this reason, it is called the *independent variable*. The *dependent variable*, y, describes measurements made at chosen values of x and is almost always inaccurate to some degree. Since the x-values are selected in advance by the observer, they are most often assumed to be known exactly. Obviously, this cannot be true if x is a real number but, usually, uncertainties in x are negligible compared to uncertainties in y. When this is not true, some very subtle complications arise.

Table 2 lists data from a recent astrophysics experiment, with measurement uncertainties explicitly recorded.[2] These data come from observations, made in 1996-1997, of comet Hale-Bopp as it approached the Sun [RAU97]. Here, the independent variable is the distance of the comet from the Sun. The unit is AU, the average distance (approximately) of the Earth from the Sun. The dependent variable is the rate of production of cyanide, CN, a decomposition product of hydrogen cyanide, HCN, with units of molecules per second divided by 10^{25}. Thus, even when Hale-Bopp was well beyond the orbit of Jupiter (5.2 AU), it was producing cyanide at a rate of $(6 \pm 3) \times 10^{25}$ molecules per second, that is, nearly 2.6 kg/s.

Table 2. Rate of Production of CN in Comet Hale-Bopp

Rate (molecules per second)/10^{25}	Distance from Sun (AU)	Uncertainty in Rate (molecules per second)/10^{25}
130	2.9	40
190	3.1	70
90	3.3	20
60	4.0	20
20	4.6	10
11	5.0	6
6	6.8	3

In this example, the uncertainties in the measurements (Table 2, column 3) are a significant fraction of the observations themselves. Establishing the value of the uncertainty for each data point and assessing the net effect of uncertainties are crucial steps in any analysis. Had Kepler's data been as poor as the data available to Copernicus, his name would be known only to historians.

The data of Table 2 are presented graphically in Figure 2. For each point, the length of the *error bar* indicates the uncertainty[3] in y. These uncertainties vary considerably and with some regularity. Here, as often happens with observations made by electronic instruments which measure a physical quantity proportional to the target variable, the uncertainty in an observation tends to increase with the magnitude of the observed value.

Qualitatively, these data suggest the hypothesis that the comet produced more and more CN as it got closer to the Sun. This would make sense since all chemical reactions go faster as the temperature increases. On the other hand, the observed rate at 2.9 AU seems too small. Did the comet simply start running out of HCN? How likely is it that the rate at 3.1 AU was really bigger than the rate at 2.9 AU? Are these values correct? Are the uncertainties correct? If the uncertainties are correct, what does this say about the validity of the hypothesis? All of these are legitimate questions.

---

[2] see file *Examples:Hale_Bopp.CN.in*

[3] In spite of its name, this bar does not indicate error. If it did, the error could be readily removed.

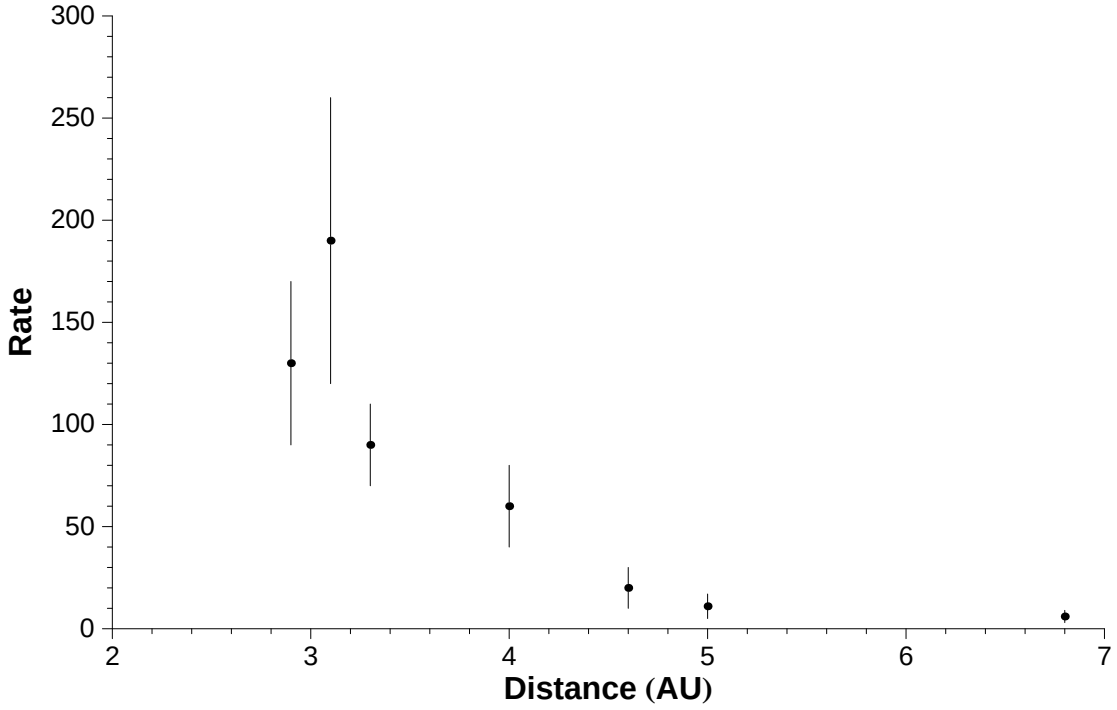

Figure 2. Hale-Bopp CN Data

Finally, consider the very "unscientific" data shown in Figure 3. This figure is a plot of the highest major-league baseball batting averages in the United States, for the years 1901-1997, as a function of time.[4]

A player's batting average is the fraction of his "official at-bats" in which he hit safely. Thus, it varies continuously from zero to one. It is fairly clear that there is a large difference between these data and those shown in Figure 1. The latter look like something from a math textbook. One gets the feeling that a daytime value could be predicted rather well from the values of its two nearest neighbors. There is no such feeling regarding the data in Figure 3. At best, it might be said that batting champions did better before World War II than afterwards. However, this is not an impressive conclusion given nearly a hundred data points.

Considering the data in Figure 3, there can be little doubt that maximum batting average is not really a function of time. Indeed, it is not a function of anything. It is a *random variable* and its values are called *random variates*, a term signifying no pretense whatever that any of these values are **individually** predictable.[5] When discussing random (stochastic) variables, the terms "independent" and "dependent" have no relevance and are not used, nor are scatter plots such as Figure 3 ever drawn except to illustrate that they are almost meaningless.

---

4 see files *Examples:BattingAvgEq.in* and *Examples:BattingAvg.in* [FAM98]

5 The qualification is crucial; it makes random data comprehensible.

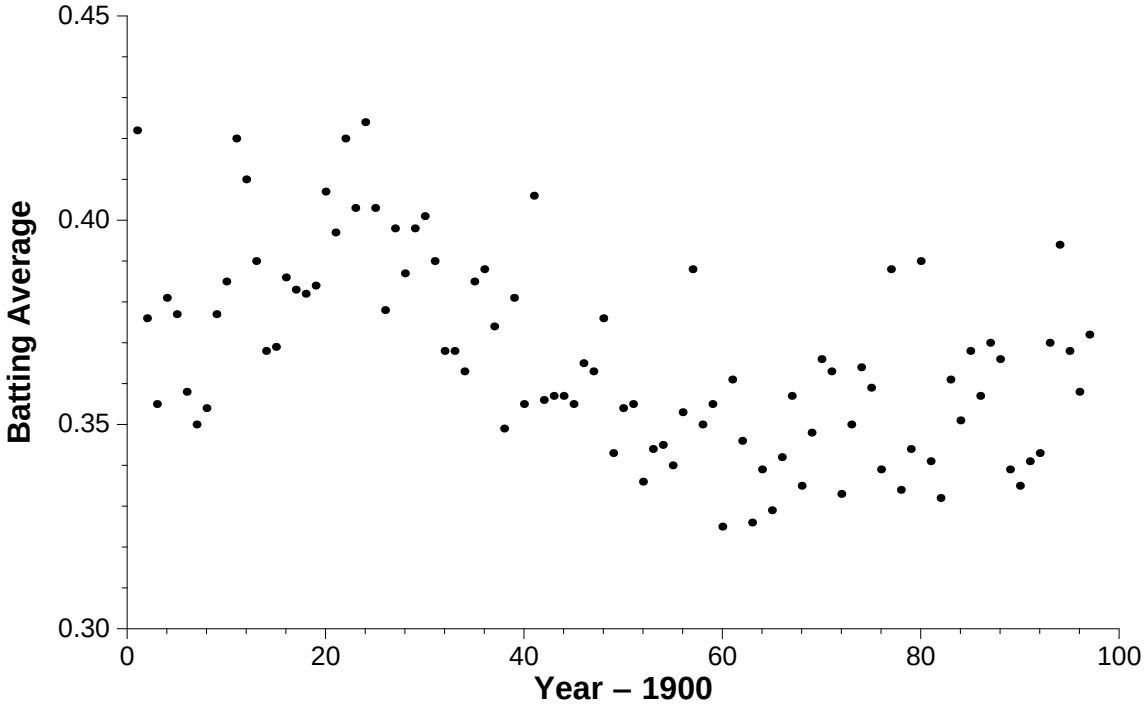

Figure 3. Annual Best Baseball Batting Average

Variables appear random for one of two reasons. Either they are inherently unpredictable, in principle, or they simply appear so to an observer who happens to be missing some vital information that would render them *deterministic* (non-random). Although deterministic processes are understandably of primary interest, random variables are actually the more common simply because that is the nature of the Universe. In fact, as the next section will describe in detail, understanding random variables is an essential prerequisite for understanding any real dataset.

Making sense of randomness is not the paradox it seems. Some of the metrics that apply to deterministic variables apply equally well to random variables. For instance, the mean and variance (or standard deviation) of these batting averages could be computed just as easily as with the daytime values in Example 1. A computer would not care where the numbers came from. No surprise then that statistical methodology may be profitably applied in both cases. Even random data have a story to tell.

Which brings us back to the point of this discussion. We have data; we seek insight and understanding. How do we go from one to the other? What's the connection?

The answer to this question was Kepler's most important discovery. Data are connected to understanding by a *model*. When the data are quantitative, the model is a mathematical model, in which case, not only does the form of the model lead directly to understanding but one may query the model to gain further information and insight.

But, first, one must have a model.

FROM DATA TO MODEL

Marshall McLuhan is widely recognized for his oft-quoted aphorism, "The medium is the message." It is certainly true in the present context. The model is the medium between data and understanding and it is, as well, the message of this tutorial. The current section presents an outline of fundamental concepts essential for initiating the first step, from data to model. The methodology for implementing these concepts is described in later sections. The second step, from model to understanding, will be left as an exercise for the reader.

As can be seen, through the examples above, data are not transparent. They do not reveal their secrets to casual observers. Yet, when acquired with due diligence, they contain useful information. The operative word is "contain." Data and information are not equivalent even though, in colloquial speech, they are treated as such.

Why not? What is there in a dataset that is not information? The answer is *error*.

$$\text{Data} = \text{Information} + \text{Error}$$

Were it not for error, every observation would be accurate, every experiment a paragon of perfection, and nearly every high school Science Fair project a fast track to the Nobel Prize. Alas, error exists. It exists not just in every dataset but in every data point in every dataset. It contaminates everything it touches and it touches everything. To make any progress, it must be identified and filtered out.

A good model does this very nicely. In its simplest, most superficial aspect, a model is a filter designed to separate data into these two components. A mathematical model is a statistical filter that not only attempts the separation but also quantifies its success in that effort.

Two Ways to Construct a Model

To construct a model, it is necessary to proceed from the known to the unknown or, at the very least, from the better known to the less well known. There are two approaches. The choice depends upon whether it is the information or the error that is better known, bearing in mind that "known" and "assumed" are not synonyms. In the first case, the model is designed to utilize the known properties of the embedded information to extract the latter and leave the error behind. This approach is commonly employed with stochastic data. Alternatively, if the error is the better known, the model is designed to operate on the error, filtering it out and leaving the information behind. This approach is nearly universal with deterministic data. In either case, the separation will be imperfect and the information still a bit erroneous.

The notion of deterministic information is commonplace and requires no further elaboration but what about the putative "stochastic information" contained in a dataset of random variables? Is it real? Likewise, can anything useful be said about error? Do random variables and error really have "properties" that one can understand? In other words, are they comprehensible?

They are indeed. The remainder of this section elucidates some properties of i) stochastic information and ii) error. Such properties are quantitative, leading directly to the identification of optimization criteria which are crucial to any modeling process.

Stochastic Information

Stochastic. Information. The juxtaposition seems almost oxymoronic. If something is stochastic (random), does that not imply the absence of information? Can accurate conclusions really be induced from random variables?

Well, yes, as a matter of fact. That a variable is stochastic signifies only that its next value is unpredictable, no matter how much one might know about its history. The data shown in Figure 3 are stochastic in this sense. If you knew the maximum batting average for every year except 1950, that still would not be enough information to tell you the correct value for the missing year. In fact, no amount of ancillary data would be sufficient. This quality is, for all practical purposes, the quintessence of *randomness*.

Yet, we have a very practical purpose here. We want to assert something definitive about stochastic information. We want to construct models for randomness.

Such models are feasible. Although future values for a random variable are unpredictable in principle, their **collective behavior** is not. Were it missing from Figure 3, the maximum batting average for 1950 could not be computed using any algorithm. However, suggested values could be considered and assessed with respect to their **probability**. For instance, one could state with confidence that the missing value is probably not less than 0.200 or greater than 0.500. One could be quite definite about the improbability of an infinite number of candidate values because any large collection of random variates will exhibit consistency in spite of the randomness of individual variates. It is a matter of experience that the Universe is consistent even when it is random.

A mathematical model for randomness is expressed as a *probability density function* (PDF). The easiest way to understand a PDF is to consider the process of computing a weighted average. For example, what is the *expectation* (average value) resulting from tossing a pair of ordinary dice if prime totals are discounted? There are six possible outcomes but they are not equally probable. Therefore, to get the correct answer, one must calculate the **weighted** average instead of just adding up the six values and dividing the total by six. This requires a set of weights, as shown in Table 3.

Table 3. Weights for Non-prime Totals for Two Dice

Value	Weight
4	3/21
6	5/21
8	5/21
9	4/21
10	3/21
12	1/21

The random total, y, has an expected value, denoted $\langle y \rangle$ or $\bar{y}$, given by Equation 3.

$$\text{Expectation of } y \equiv \langle y \rangle = \sum_{i=1}^{6} w_i \, y_i \qquad 3.$$

The true expectation, 7.62, is quite different from the unweighted average, 8.17.

The set of weights in Table 3 constitutes a *discrete* PDF, meaning that there are a **countable** number of possible values, each with its own weight (density). Given a discrete PDF, f(y), for the random variable, y, any arbitrary function of y, g(y), will have an expectation computed as shown in Equation 4. Equation 3 is just a special case of Equation 4.

$$\langle g(y) \rangle = \sum_{\text{all } i} f(y_i) \, g(y_i) \, \Delta y \qquad 4.$$

where Δy is equal to the *binwidth*, defined below in Example S1.

In the *continuous* case, neighboring values are separated by the infinitesimal binwidth, dy. Nevertheless, a continuous PDF has exactly the same meaning as a discrete PDF and, thus, the continuous analogue of Equation 4 is Equation 5.

$$\langle g(y) \rangle = \int_{-\infty}^{\infty} f(y) \, g(y) \, dy \qquad 5.$$

For example, suppose $f(y) = e^{-y}$ and $y \geq 0$. Then, the expectation of $g(y) = \sqrt{y}$ is

$$\langle \sqrt{y} \rangle = \int_{0}^{\infty} e^{-y} \sqrt{y} \, dy = \frac{\sqrt{\pi}}{2} \qquad 6.$$

The product $(f(y) \, \Delta y)$ or $(f(y) \, dy)$ equals the probability of a given y. Probability is a dimensionless quantity. That is, it is a pure number with no units. Consequently, the units of a PDF must always be the reciprocal of the units, if any, of y.[6] It is for this reason that the function f(y) is referred to as a "density" function.

Example S1—Batting Averages

The dataset shown in Figure 3 will be the first example. The collective behavior of these data can be summarized numerically and graphically. For instance, their mean (expectation) is 0.367 and their variance is 0.0005596 [st. dev. = 0.0237]. Any statistics textbook will list several additional metrics that could be used to characterize this set of 97 values. For a graphical depiction, the device most commonly employed is the *histogram*, constructed by grouping the variates into numerical intervals of equal width and plotting these *bins* against their *frequencies*, the number of variates in the respective bins. These frequencies may be converted to probabilities (*normalized*) by dividing each by the sample size and then converted to probability densities by further dividing by the binwidth (here, 0.011). When all of this is done with the batting-average data, we get the histogram shown in Figure 4 (gray boxes).

---

6 a useful check for complicated density formulas (see Appendix A)

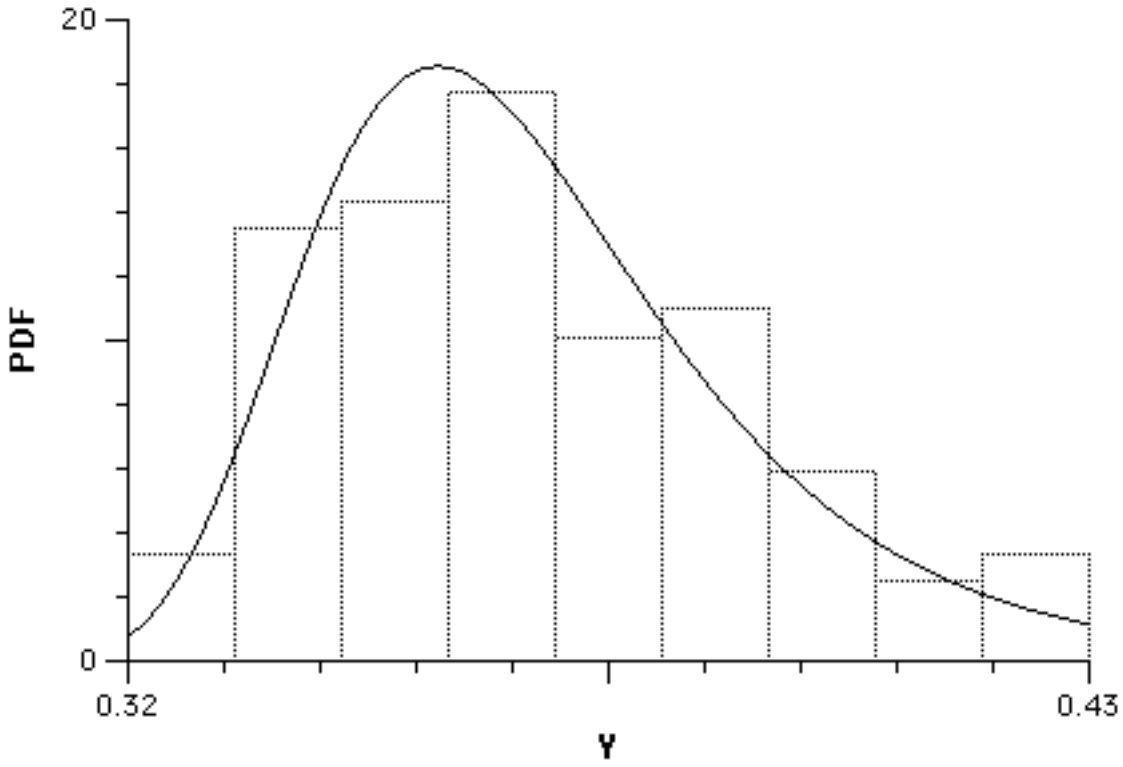

Figure 4. Batting-Average Data (Histogram and PDF)

The overall shape of a histogram is intended to describe how realizable values of a random variable are scattered across the infinite landscape of conceivable locations. The way it works is that the histogram is constructed to have a total **area** = 1. Therefore, the probability that any random variate falls in a given bin is equal to the area of the corresponding rectangle. This scattering of values/probabilities is referred to as the *distribution* of the random variable. Note that the Y-axis measures probability density, not probability. Given the computational process just outlined, and recalling Equations 4 and 5, measurements along this axis have units of 1/u, where u is the unit of the random variable (e.g., inches, grams, etc.)

A continuous density function for these data is also shown in Figure 4 (solid line). The corresponding analytical expression, given in Equation 7, is called the Gumbel distribution.[7] Like the histogram, the total area under this curve, including the tails (not shown), equals 1.

$$\text{PDF} = \frac{1}{B} \exp\left(\frac{A-y}{B}\right) \exp\left(-\exp\left(\frac{A-y}{B}\right)\right) \qquad 7.$$

where, for this example, A = 0.3555 and B = 0.01984.

[7] see pg. A-49

We see that Equation 7 does have the proper units. Parameter A has the same units as the random variable, y. Since any exponent is necessarily dimensionless, parameter B must have the same units as well. Hence, the entire expression has units of 1/u. Of course, these particular data happen to be dimensionless already, but variates described by a PDF usually do have units of some sort.

The Gumbel distribution is one example of a continuous distribution. [There are many others described in Appendix A.] If f(y) is the PDF, the probability of any value of y would equal f(y)*dy which, since dy is vanishingly small, is zero. Indeed, intuition demands that the probability of randomly picking any given value from a continuum must be zero. However, the probability of picking a value in a finite range [a, b] may be greater than zero (Equation 8).

$$\text{Prob}\left(a \le y \le b\right) = \int_a^b f\left(y\right) dy \qquad 8.$$

The integral (or sum) of a density function, from minus infinity to some chosen value, y, is referred to as the *cumulative distribution function* (CDF), usually denoted F(y).[8] For this example, it is plotted in Figure 5. In this figure, the solid line is the theoretical CDF and the gray line the empirical CDF. The CDF-axis gives the probability that a randomly selected variate will be less than or equal to b (Equation 9).

$$\text{Prob}\left(y \le b\right) = F\left(b\right) = \int_{-\infty}^b f\left(y\right) dy \qquad 9.$$

Some of the reasons for choosing the Gumbel distribution for these data are discussed in its entry in Appendix A. For now, it is sufficient to appreciate that any sort of "theoretical" expression would be considered relevant. This supports the conclusion that useful statements can be made about stochastic data. However, it does not explain where these parameter values came from. They are another matter entirely.

One of the reasons for using the values of A and B given above is that they produce a curve that encloses an area with approximately the same shape and size as the empirical histogram. There are much better reasons but their explication requires some preliminary concepts.

The first is the *likelihood* of a dataset. The term means just what it says. It indicates how "likely" this set of observations would be had they been selected randomly from the given PDF. **If the variates are independent**, the likelihood is also a measure of their *joint probability*.[9] For any PDF, f(y), the likelihood of a sample of N variates is defined by Equation 10.

$$\text{Likelihood} \equiv \prod_{k=1}^{N} f\left(y_k\right) \qquad 10.$$

---

[8] The term "distribution" is sometimes taken to mean the cumulative distribution.

[9] Two random quantities, x and xx, are independent if and only if their joint (simultaneous) probability, Prob(x AND xx), is **always** equal to Prob(x)*Prob(xx).

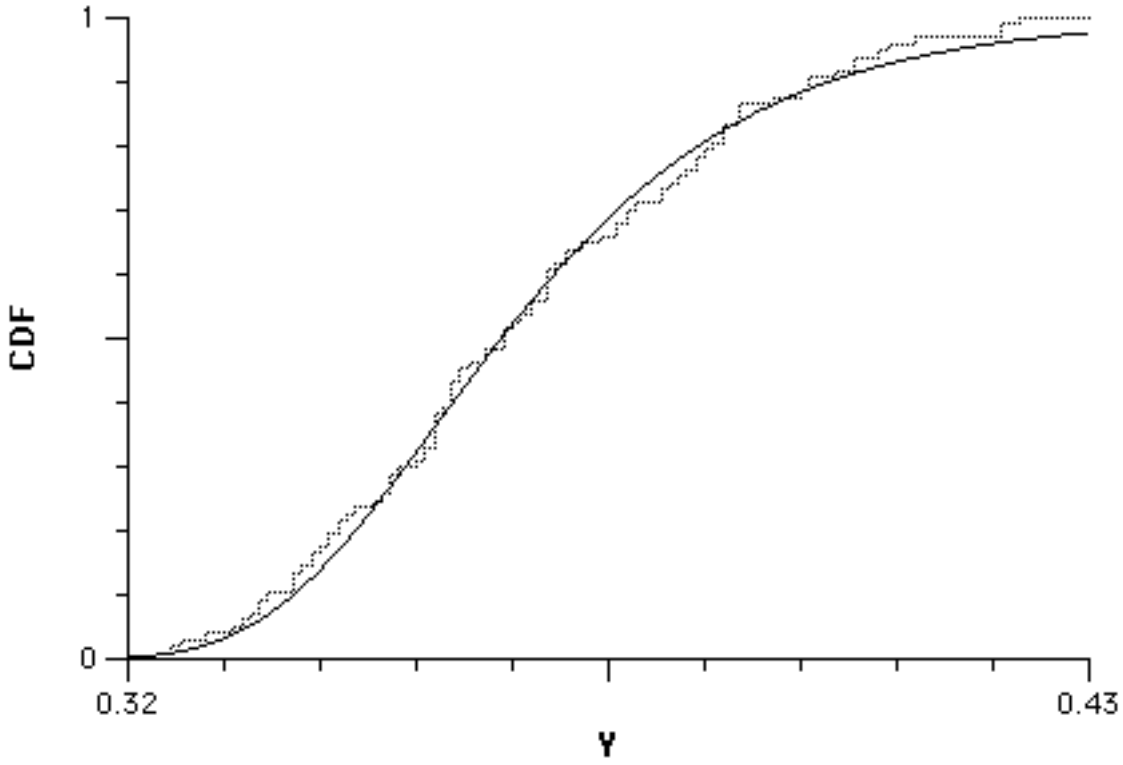

Figure 5. Batting-Average Data (CDF)

The second concept is that of *maximum likelihood* (ML). This notion is also very intuitive. Given a form for the density function, the ML parameters are those parameters which make the likelihood as large as possible. In other words, any other set of parameters would make the observed dataset less likely to have been observed. Since the dataset was observed, the ML parameters are generally considered to be optimal.[10] However, this does not preclude the use of some alternate criterion should circumstances warrant.

The values given above for A and B are the ML parameters. This is an excellent reason for choosing them to represent the data and it is no coincidence that the theoretical curve in Figure 4 matches the histogram so well. Had there been a thousand data points, with proportionately narrower histogram bins, the match would have been even better.

Thus, a Gumbel distribution, with these ML parameters, is a model for the data shown in Figure 3. It summarizes and characterizes the stochastic information contained in these data. Not only does it describe the probability density of the data, but one could query the model. One could ask the same questions of the model that could be asked of the data. For example, "What are the average and standard deviation of these data?" or "What is the probability that next year's batting champion will have a batting average within five percent of the historical

[10] Their uniqueness is usually taken for granted.

mean?" The answers to such questions may be computed from the model without looking at the data at all.

For instance, the last question is answered in Equation 11.[7]

$$\text{Probability} = \int_{0.95 * \text{mean}}^{1.05 * \text{mean}} f\left(y\right) dy = F\left(0.385\right) - F\left(0.349\right) = 0.55 \qquad 11.$$

As a check, note that Equation 11 predicts that 53 of the values in Figure 3 are within the given range. The data show 52 in this range.

One could even ask questions of the model that one could not ask of the data. For instance, "What is the probability that next year's batting champion will have a batting average greater than the maximum in the dataset (0.424)?" Obviously, there is no such number in the historical data and that suggests an answer of zero. However, this is clearly incorrect. Sooner or later, someone is bound to beat 0.424 provided that major-league baseball continues to be played.

It is easy to pose a theoretical question like this to a model and, **if the model is good**, to obtain a correct answer. This model (Equation 7) gives an answer of three percent, suggesting that we are already overdue for such a feat.

Whether or not a model is good (valid) is a question yet to be addressed.

Example S2—Rolling Dice

The Gumbel distribution was chosen to model the batting-average data because it is known to be appropriate, in many cases, for samples of continuous-distribution maxima (so-called *record values*). However, there is no *a priori* guarantee that it is valid for these particular data. Occasionally, there is enough known about some type of stochastic information that theory alone is sufficient to specify one distribution to the exclusion of all others and, perhaps, even specify the parameters as well. The second stochastic example illustrates this situation *via* the following experiment.

Step 1
Think of a number, a whole number, from one to six.
Step 2
Roll a standard, cubical die repeatedly until your chosen number appears three times.
Step 3
Record the number of rolls, N_r, required.

After only a few semesters of advanced mathematics courses, it would be relatively easy to prove that the random variable, N_r, is ~NegativeBinomial(1/6, 3).[11] This assertion could be tested by performing the experiment a very large number of times and comparing the theoretical distribution to the empirical data.

[11] see pg. A-81; the ~ is read "(is) distributed as"

Carrying out this exercise by hand would be a bit tedious. Fortunately, it is a simple matter to *simulate* the experiment on a computer. Such simulations are very common. Simulated results for a 1,000-fold repetition of this experiment are shown in Figure 6.[12]

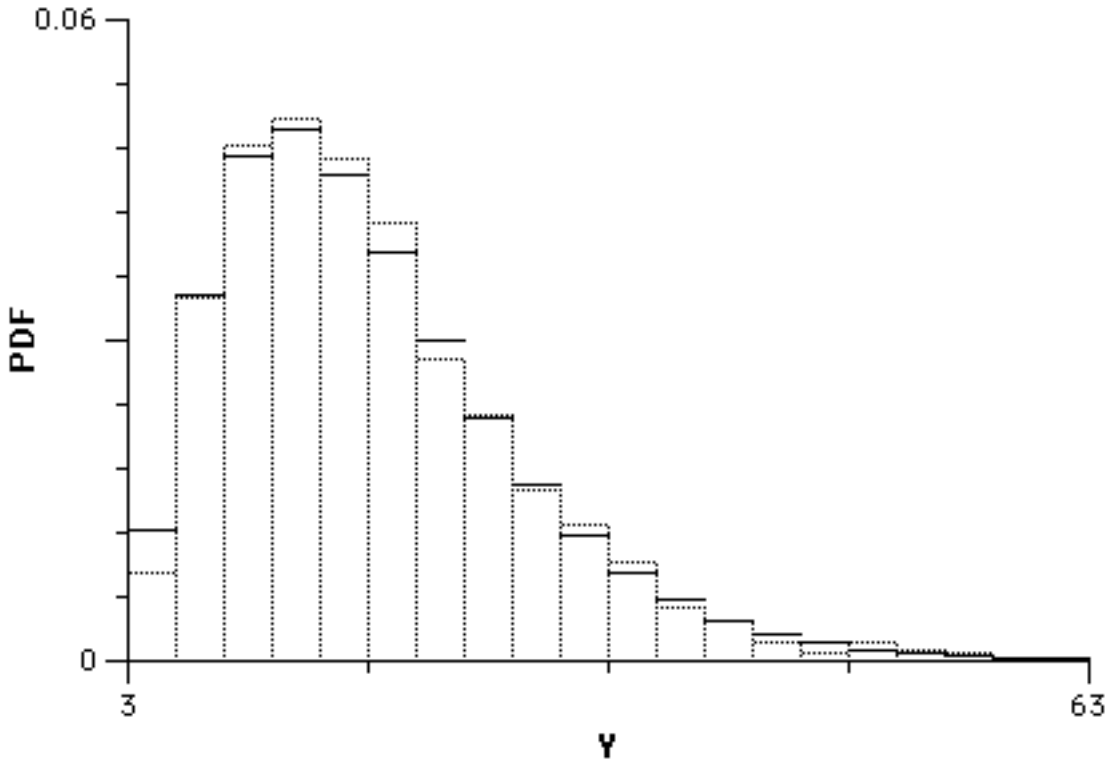

Figure 6. Rolls3 Data (Histogram and PDF)

The NegativeBinomial model shown in Figure 6 (solid lines) is another example of a discrete PDF. Typically, this means that the random variable can take on only integer values. That is the case here since N_r is a count of the number of rolls required.

In computing the model shown, the first parameter was not fixed at its theoretical value, 1/6. Instead, it was estimated from the data. The ML estimate = 0.1691, very close to the theoretical 0.1667. It would have been even closer had the sample size been much larger than 1,000. Likewise, the mean predicted by this model = 17.74, very close to the theoretical mean, 18. The second parameter, 3, was assumed given although it, too, could have been considered unknown and estimated from the data.

The observed data (gray histogram) may be compared to the estimated, ML model using the *Chi-square* statistic, defined in Equation 12.

[12] see file *Examples:Rolls3.in*

$$\text{Chi–square} \equiv \chi^2 = \sum_{\text{all k}} \frac{\left(o_k - e_k\right)^2}{e_k} \qquad 12.$$

where k includes all possible values of the random variable and where o and e are the observed and expected frequencies, respectively.

For this example, Chi-square = 65.235. Again, the question of whether this amount of discrepancy between theory and experiment is "acceptable" is a matter to be discussed below.

Example S3—Normality

We would be remiss in our illustrations of stochastic information were we to omit the most famous of all stochastic models, the Normal (Gaussian) distribution. The Normal distribution is a continuous distribution and it arises naturally as a result of the *Central Limit Theorem*. In plain English, this theorem states that, provided they exist, averages (means) of any random variable or combination of random variables tend to be ~Normal(A, B), where A is the sample mean and B is the unbiased estimate of the population standard deviation. As usual, with random data, the phrase "tend to" indicates that the validity of this Gaussian model for random means increases as the sample size increases.

In Example S2, a thousand experiments were tabulated and individual outcomes recorded. A sample size of 1,000 is certainly large enough to illustrate the Central Limit Theorem. Our final example will, therefore, be a 1,000-fold replication of Example S2, with one alteration. Instead of recording 1,000 values of N_r for each replicate, only their average, N_{avg}, will be recorded. This new experiment will thus provide 1,000 averages[13] which, according to the Central Limit Theorem, should be normally distributed.

As before, the experiment was carried out in simulation. The observed results and the ML model are shown in Figure 7. The estimated values for A and B are 17.991 and 0.2926, respectively. With a Gaussian model, the ML parameter estimates may be computed directly from the data so one cannot ask how well they match the observed values.

A somewhat better test of the model would be to pick a statistic not formally included in the analytical form of the distribution. One such statistic is the interquartile range, that is, the range included by the middle half of the sorted data. In this example, the model predicts a range of [17.79, 18.19][14] while the data show a range of [17.81, 18.19].

The ideal test would be one that was independent of the exact form of the density function. For continuous distributions, the most common such test is the *Kolmogorov-Smirnov* (K-S) statistic. The K-S statistic is computed by comparing the empirical CDF to the theoretical CDF. For example S3, these two are shown in Figure 8. The K-S statistic is simply the magnitude (absolute value) of the maximum discrepancy between the two, measured along the CDF-axis. Here, the K-S statistic = 0.0220.

---

[13] see file *Examples:Rolls3avg.in*

[14] see pg. A-85

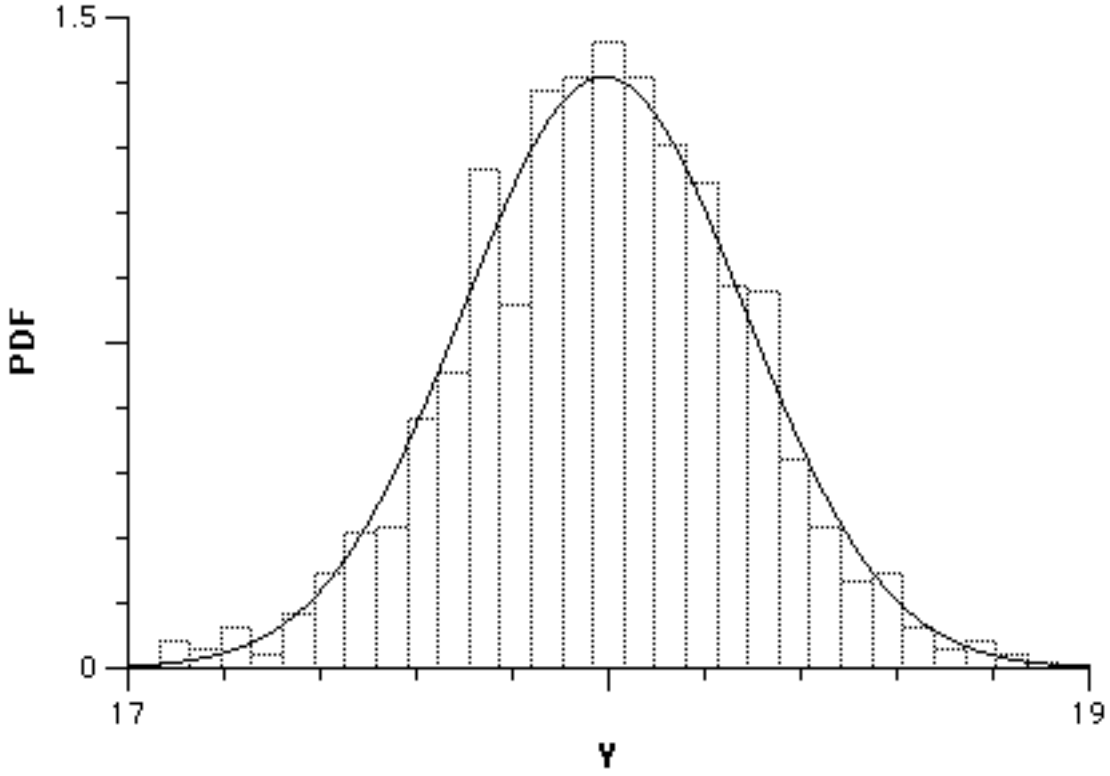

Figure 7. Rolls3avg Data (Histogram and PDF)

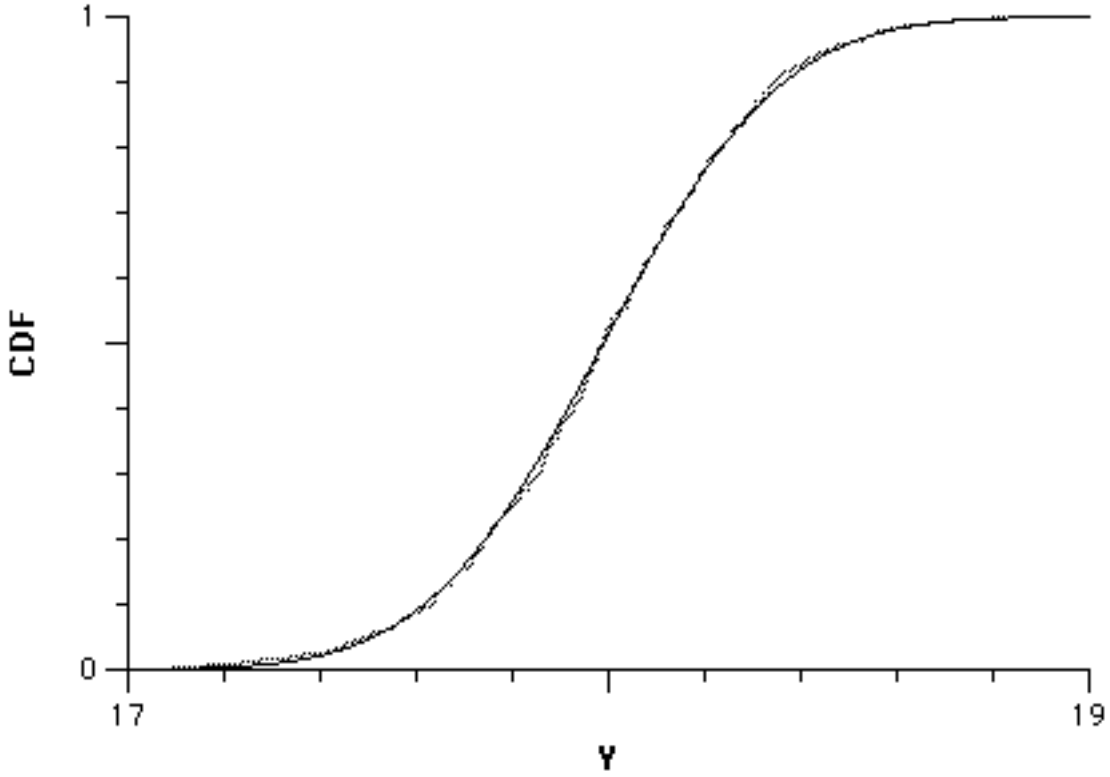

Figure 8. Rolls3avg Data (CDF)

Once again, the question naturally arises as to whether this result is acceptable and, once again, we defer the question until later.

Stochastic Information—Summary

In examples S1 to S3, we have endeavored to show that the term "stochastic information" is not the contradiction it would appear to be. It should now be clear that random variables are collectively, if not individually, predictable to a useful extent and that much can be said, with accuracy, about such variables.

Since much can be said, it follows that it should be possible to use this information to filter out some of the error found in stochastic data. Just because variates are random, it does not mean that they are errorless. The data of Example S1, for instance, appeared to be described by a Gumbel distribution. However, as with all stochastic datasets, there were discrepancies between model and data. How much of this was due to inherent variation and how much was due to error was not discussed. It will be.

Finally, we have described three metrics: the ML criterion, the Chi-square statistic, and the K-S statistic. These three are valid and useful concepts for the description of stochastic data and it will soon become evident that all may be utilized as optimization criteria as well.

First, however, we discuss deterministic information and the kinds of error commonly associated with it.

Deterministic Information and Error

If "stochastic information" seems an oxymoron, "deterministic information" would, for similar reasons, appear to be a tautology. Neither is true and, in this discussion, the latter term will refer to any information gleaned from non-stochastic sources, typically experimentation. Since these sources are prone to error, it is essential to characterize that error in order to attempt to compensate for it. This section will focus on the kinds of errors found in deterministic data.

The amount of error in a dataset, or even a data point, is generally unknown. The reason, of course, is that error is largely random. We have just seen, however, that randomness does not imply total ignorance. To the contrary, modeling of deterministic data usually assumes that something valid is known about the error component of the data, providing a handle with which to filter it out. In fact, whenever data of a given type are investigated over an extended period of time, it is not uncommon for associated errors to become as well characterized as the information itself.

However, for the purposes of this tutorial, our goals are much less specific. We shall limit our illustrations to a few kinds of error and say something, as well, about error in general. As in the last section, we seek quantitative metrics that may be employed as optimization criteria.

Components of Error

Empirical datasets, almost without exception, are contaminated with error. Even the data shown in Table 1, which few would dispute, are recorded with limited precision. Therefore, they are subject, at the very least, to *quantization* error. A dataset without error might be imagined but hardly ever demonstrated.

Not only does a dataset, as a whole, contain an unknown amount of error but every data point in it contains some unknown fraction of that total error. Consequently, the amount of information in a dataset is unknown as well. Each data point makes some contribution to the total amount of information but there is no way to tell how much just by looking. Information and error have the same units and may be physically indistinguishable.

Error comprises some combination of *bias* and *noise*. A bias is a systematic deviation from the truth due either to a deterministic deviation, consistently applied, or to a random deviation with a non-zero mean. The former is seen, for instance, when your bathroom scale is not properly zeroed in. In this case, it will report your apparent weight as too high or too low, depending on the sign of the bias in the zero-point. The latter type of bias is exemplified in the case of an astronomer who records the image of a distant galaxy by counting the photons that impinge on the pixels of a light-sensitive array. Inevitably, each pixel is subject to a counting error. Such errors are typically random variables ~Poisson(A) and have a mean, $A > 0$.[15]

Noise refers either to random, zero-mean errors or to the residual portion of random, biased errors once the bias has been removed. For example, undergraduate chemistry students, performing their first quantitative, organic analysis usually get experimental results exhibiting lots of errors. These errors are largely random but often have a significant negative

---

[15] see pg. A-101

bias because, in trying to isolate the target substance, some is lost. Quantifying the remainder is a process usually subject to zero-mean, random errors, i.e., noise.

Most often, total error represents the combined effect of several sources of different kinds of error, including biases which can sometimes be identified and removed, as well as noise. To the degree that error is random, only its average effects can be characterized and there is, of course, no guarantee that a given sample will be average.

As before, we shall eschew theoretical discussions in favor of concrete examples. There is ample literature available to anyone wishing to pursue this subject in depth. The examples below all relate to deterministic data.

Example D1—Quantization Error

The daytimes in Table 1 provide an illustration of one kind of error and, at the same time, of the utility of the variance.

One of the most useful properties of the variance statistic is its *additivity*. Whenever any given variance, V, is due to the combined effect of k **independent** sources of variation, V will equal the sum of the individual variances of these k sources. [Note the recursive nature of this property.]

The variance of the daytimes listed in Table 1 = TSS/N = 836,510/43 = 19,453.72 min^2. As noted above, a portion of this variance derives from the fact that these values, having been rounded to the nearest minute, have limited precision. This round-off error has nothing to do with the mechanism responsible for the secular variation in daytime. Hence, the variance due to rounding (quantization) is independent of all remaining variance, so it can be factored out.

It is easy to demonstrate that quantization error is a random variable ~Uniform[–B, B], where B is one-half the unit of quantization.[16] As such, the variance due to this single source of error, Q, is given by Equation 13. For these data, the unit of quantization is one minute. Therefore, Q = 0.08 min^2, only 0.0004 percent of the total variance of the data.

$$\text{Quantization variance} \equiv Q = \frac{\left(\text{unit of quantization}\right)^2}{12} \qquad 13.$$

Continually citing squared quantities (with squared units) is inconvenient and the standard deviation is quoted more often than is the variance. Unfortunately, standard deviations are **not** additive. In this example, the "standard deviation of quantization error" is 0.29 min (about 17 seconds). Thus, although this quantization error is negligible compared to the overall variation of the data, it is clearly not negligible when compared to the recorded precision.

If quantization were the only source of error for this dataset, the deterministic information present would constitute 99.9996 percent of the observed variance. It would then be left to the chosen model to "explain" this information. Note that a deterministic model is not expected to

16 see pg. A-113

explain error. That task is delegated to a separate error model, implicitly specified as part of the optimization procedure.

Finally, the variance statistic is evidently useful not just for computational purposes but, as seen here, as a measure of information. In other words, were the errorless component of the total variance equal to zero, there would be nothing to explain. In Example D2, we utilize this aspect of the variance to develop a quantitative metric for deterministic information explained.

Example D2—Kepler's First Law

In a deterministic model, the value of the independent variable is supposed to be sufficient to predict the value of the dependent variable to an accuracy matching that of the input. If the model is important enough, it is often referred to as a *law*. The implied reverence generally signifies that something of unusual significance has been discovered about the real world and summarized in the model so designated. Today, in the physical sciences, laws are particularly precise, as are their parameters. In fields related to physics, for instance, it is not uncommon to find model parameters accurate to more than nine significant figures (i.e., one part per billion). This mathematical approach to understanding we owe, in large part, to the efforts of Johannes Kepler (1571-1630) and Galileo Galilei (1564-1642).

This example was presented by Kepler as an illustration of his First Law, arguably the first correct mathematical model correctly devised. The law states that each planet revolves about the Sun in an ellipse with the Sun at one focus.

Figure 9 depicts an ellipse, together with some of its more important features. Like any ellipse, its size and shape may be completely specified by two numbers. Usually, one of these is the length of the *semi-major axis*, OP, and the other is the *eccentricity*, which is the ratio of two lengths, OS/OP. The ellipse in the figure has an eccentricity of 0.5 so one focus, S, is halfway between the center, O, and the circumference at P; the other focus, not shown, is the mirror image of S through the center.

Kepler spent much of his life investigating the orbit of Mars. This orbit has an eccentricity of only 0.09, so it is much more circular than the ellipse shown in Figure 9 as, indeed, are the orbits of all planets in the Solar System, except Mercury and Pluto. The point P represents the point closest to the Sun, S, and is called *perihelion*; the point farthest from the Sun, A, is called *aphelion*. A chord, SG, drawn through a focus and parallel to the *semi-minor axis*, OR, is called the *semi-latus rectum*. The latter, plus the eccentricity, are the only parameters required when an ellipse is described in polar coordinates (Equation 14).

$$\text{radius} = \frac{A}{1 + B \cos(\nu)} \qquad 14.$$

where the origin is at S, from which the radius (e.g., SM) is measured, and where the angle, ν, measured from perihelion in the direction of motion, is called the *true anomaly*.

In Equation 14, parameter A is the length of the semi-latus rectum and B is the eccentricity. This equation is actually a general equation describing all of the conic sections—circle, ellipse, parabola, or hyperbola, depending upon the eccentricity (see Table 4). The independent

variable is true anomaly, from which the dependent variable, the heliocentric radius, may be computed/predicted **if the model is valid**.

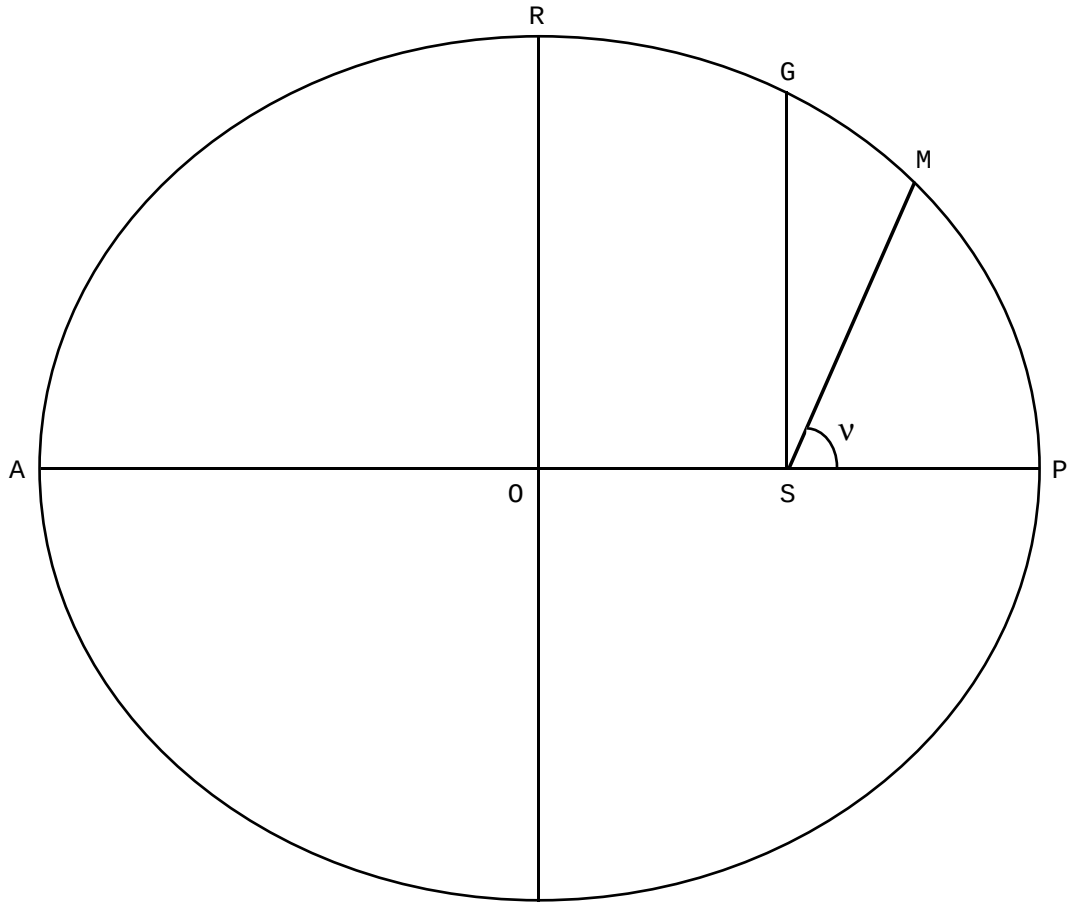

Figure 9. Ellipse (eccentricity = 0.5)

Table 4. Eccentricities of Conic Sections

Eccentricity	Shape
0	Circle
(0, 1)	Ellipse
1	Parabola
> 1	Hyperbola

With these preliminaries, we may now examine Kepler's data, shown in Table 5 in modern units, along with the actual values for the times indicated [KEP09, STA97].[17]

[17] see files *Examples:Kepler_Mars.in* and *Examples:True_Mars.in* Note that these data do not comprise a single Martian year; they were observed during a sequence of oppositions.

Table 5. Orbital Data for Mars

Time (JD – 2300000)[18]	Radius (AU) Kepler	Radius (AU) Predicted	Radius (AU) Actual	True Anomaly (radians) Kepler	True Anomaly (radians) Actual
–789.86	1.58852	1.58853	1.58910	2.13027	2.12831
–757.18	1.62104	1.62100	1.62165	2.39837	2.39653
–753.19	1.62443	1.62443	1.62509	2.43037	2.42854
–726.27	1.64421	1.64423	1.64491	2.64327	2.64152
–30.95	1.64907	1.64907	1.64984	2.70815	2.70696
2.84	1.66210	1.66210	1.66288	2.96844	2.96732
13.74	1.66400	1.66396	1.66473	3.05176	3.05065
49.91	1.66170	1.66171	1.66241	3.32781	3.32675
734.17	1.66232	1.66233	1.66356	3.30693	3.30640
772.03	1.64737	1.64738	1.64812	3.59867	3.59818
777.95	1.64382	1.64383	1.64456	3.64488	3.64440
819.86	1.61027	1.61028	1.61083	3.97910	3.97865
1507.15	1.61000	1.61000	1.61059	3.98145	3.98157
1542.94	1.57141	1.57141	1.57186	4.28018	4.28035
1544.97	1.56900	1.56900	1.56944	4.29762	4.29779
1565.94	1.54326	1.54327	1.54210	4.48063	4.48083
2303.05	1.47891	1.47889	1.47886	4.94472	4.94565
2326.98	1.44981	1.44981	1.44969	5.18070	5.18169
2330.96	1.44526	1.44525	1.44512	5.22084	5.22184
2348.90	1.42608	1.42608	1.42589	5.40487	5.40591
3103.05	1.38376	1.38377	1.38332	6.12878	6.13067
3134.98	1.38463	1.38467	1.38431	0.198457	0.200358
3141.90	1.38682	1.38677	1.38643	0.274599	0.276497
3176.80	1.40697	1.40694	1.40676	0.653406	0.655260
3891.17	1.43222	1.43225	1.43206	0.940685	0.943206
3930.98	1.47890	1.47888	1.47896	1.33840	1.34079
3937.97	1.48773	1.48776	1.48789	1.40552	1.40788
3982.80	1.54539	1.54541	1.54583	1.81763	1.81986

Since Kepler intended these data to be illustrative of his law of ellipses, it is reasonable to ask whether they are, in fact, described by the model of Equation 14. The graph in Figure 10, where X is true anomaly and Y is radius, makes it qualitatively apparent that they are.[19]

This assessment may be quantified by **hypothesizing** that the information explained by this model and the residual error constitute independent sources of variation. If this is true, then the total variance of the dataset is the sum of the variances for each of these components and it would then be a trivial matter to compare the fraction of the variance/information

[18] JD (Julian date) ≡ interval, in days, from Greenwich noon, 1 January 4713 B.C.

[19] The ellipse, in this (optimized) model, has A = 1.51043 AU and B = 0.0926388.

explained by this model to the total variance in the dataset. A good model is one that leaves relatively little unexplained.

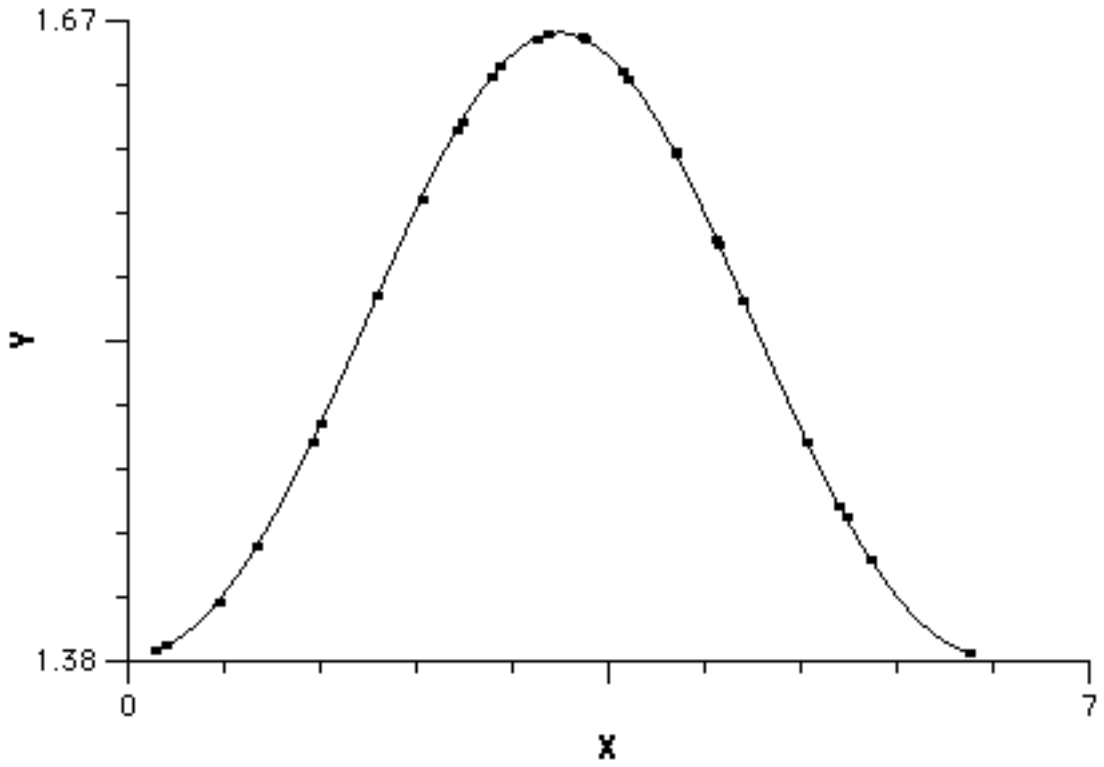

Figure 10. Kepler's Data and Model

Since the number of points is constant throughout this example, we may substitute sums of squares for the respective variances. The information not explained by the model, plus the variance due to any errors, is then equal to the *error-sum-of-squares*, ESS, often called the *residual-sum-of-squares*, defined in Equation 15.

$$\text{Error-sum-of-squares} \equiv \text{ESS} = \sum_{i=1}^{N} \left(\hat{y}_i - y_i\right)^2 \qquad 15.$$

where $\hat{y}$ is the value predicted by the model.

Therefore, the fraction of the total variance in the dataset that is explained by a given model is equal to the statistic *R-squared*, defined in Equation 16.

$$\text{R-squared} = 1 - \frac{\text{ESS}}{\text{TSS}} \qquad 16.$$

Kepler's data, plus the predictions of the optimum model (Table 5, column 3), produce the following results (computed to six significant figures):

TSS = 0.273586 AU^2
ESS = 1.257e–08 AU^2
R-squared = 1.00000

In other words, by this statistic/criterion, the model described in Equation 14 and Figure 10 explains all of the deterministic information contained in Kepler's dataset to a precision of one part in 100,000.

This result is not quite perfection. Since ESS > 0, there remains something, true deviations or error, yet to be explained.[20] Also, the 28 radii listed by Kepler are not really as accurate as his values indicate. Nevertheless, this degree of success is impressive. There are thousands of sociologists, psychologists, and economists who would be more than delighted with results of such precision.

We shall see later how Kepler's model was optimized to find parameters yielding this value for R-squared.

Example D3—World Track Records

As a measure of information and/or error, the variance statistic reigns supreme in most of contemporary analysis of deterministic data. However, intuition suggests that the "best" deterministic model is the one that minimizes the *average deviation*, defined in Equation 17. When graphed, this model gives the curve that is as close as possible to all of the data points simultaneously, using a statistic with the same units as the dependent variable.

$$\text{Average Deviation} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right| \qquad 17.$$

where $\hat{y}$ is the value predicted by the model.

To illustrate this metric, consider the data in Table 6.[21] This table lists 14 distances for which world track records (for men) are recognized. The corresponding average speed attained during each record run is listed as well [YOU97]. The task is to model this speed as a function of distance.

Not surprisingly, the average speed decreases sharply, at first, but then seems to level off somewhat. Many analytical forms would be suitable but we shall examine the power law given in Equation 18 (where y is speed and x is distance).

$$y = A\,x^B + C \qquad 18.$$

[20] That this "something" is relatively small does not imply that it is spurious or insignificant.

[21] see file *Examples:Track.in*

Table 6. World Track Records (Men)

Speed (m s^{-1})	Distance (m)
10.152	100
10.173	200
9.240	400
7.864	800
7.565	1,000
7.233	1,500
7.172	1,609.344 (mile)
6.947	2,000
6.740	3,000
6.449	5,000
6.236	10,000
5.855	20,000
5.636	25,000
5.598	30,000

The model of this form with the minimum average deviation has the parameter values given below. Its graph is shown in Figure 11.

A = 18.35
B = –0.2380
C = 4.020

The value of the average deviation resulting from this model is 0.163 meters per second; the worst deviation is –0.954 meters per second (point #2). According to the R-squared metric, this model explains 94.691 percent of the total information in the dataset.

Had we, instead, used the R-squared criterion to optimize the model, we would then have obtained the following parameters:

A = 20.36
B = –0.2369
C = 3.785

The latter model would have explained 96.634 percent of the information in the dataset but with an average deviation of 0.203 meters per second and a worst deviation of –0.582 meters per second (point #2). Evidently, just as there is more than one conceivable model for these data, there is also more than one way to optimize that model.

The minimum deviation is often used with data thought to contain one or more *outliers*, data points with unusually large deviations from the putative model. Minimizing the average deviation is more *robust* than minimizing ESS, signifying, among other things, that the computed model is not as sensitive to the presence of extreme values as it is when apparent deviations are squared. The literature on robust estimation is quite large and several good techniques have been developed over the years.

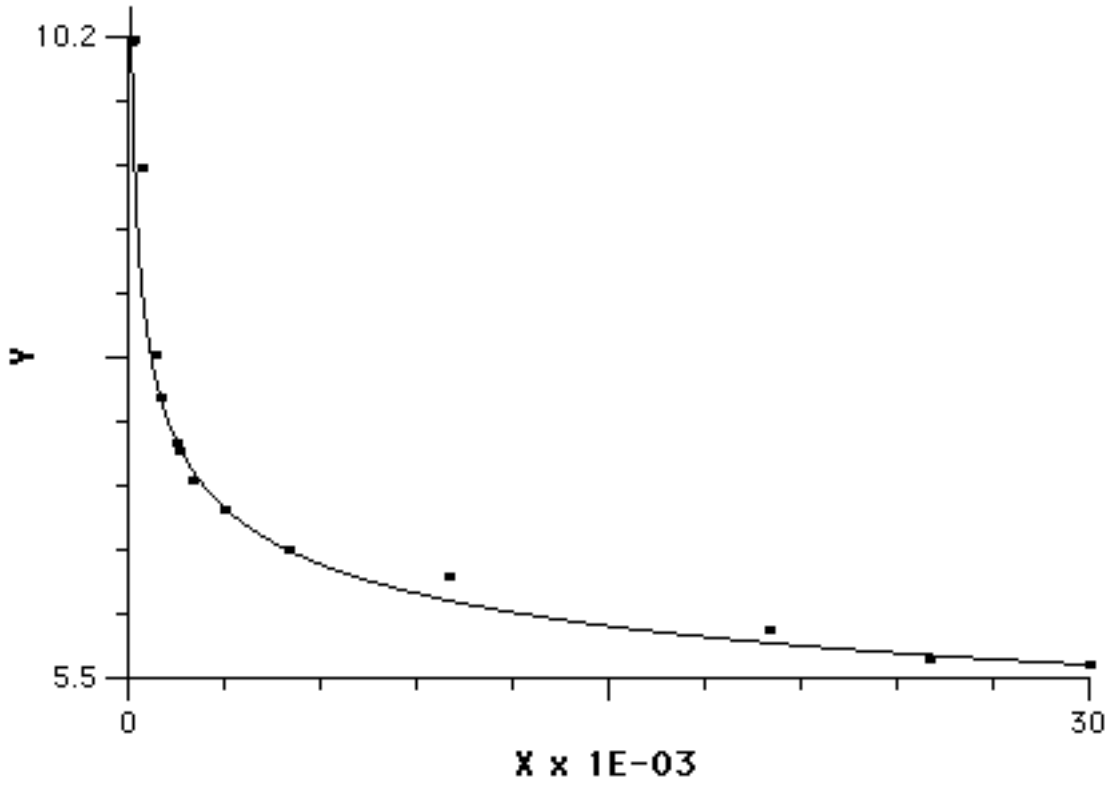

Figure 11. World Track Records

Deterministic Information—Summary

In this section, we have described two metrics for deterministic modeling, giving different answers. Is the choice a matter of indifference or is one better than the other?

To anyone trying to understand the real world, there can be only one "best" answer because there is only one Universe. Nature does not use one set of rules on weekdays and another on weekends. To be credible, a model must eschew arbitrariness to the greatest extent possible. Moreover, it must be completely consistent with whatever is already known about the data it describes, including even aspects of the data that are not the subject of the model.

As noted earlier, any deterministic model should be optimized utilizing whatever is known about the errors in the data. In Examples D2 and D3, this concern was not addressed. So far, we have considered only the fraction of information explained and average deviation (error), without considering specific properties of the residual variance. Finally, in discussing these two models, we did not relate much of what we said about stochastic modeling, especially the powerful concept of maximum likelihood, to the fact that errors in data are largely random.

In the next section we shall put everything together. Our efforts will be rewarded with a set of valid, intuitive, and well-defined criteria for computing optimum model parameters.

FINDING OPTIMUM PARAMETERS

Data, as we have seen, come in two varieties: stochastic and deterministic. Each of these may be modeled and that model characterized by one of a number of metrics (statistics) relating something about the quality of the model. Quality implies descriptive or explanatory power, proof of which is manifested in predictive capability. All of this presupposes, of course, that models exhibiting these desiderata can actually be found.

This section clarifies some relationships among the five metrics described earlier. In doing so, a general technique for finding optimum parameters will be cited. It will be apparent that this technique may be readily implemented as a computational algorithm.

Five Criteria

The five metrics/statistics that we shall utilize as modeling criteria are listed in Table 7. The two primary criteria are shown in **bold** and their preeminence will be discussed.

Table 7. Modeling Criteria

Criterion	Applicability
Maximum likelihood	All random variates
Minimum K-S statistic	Continuous random variates
Minimum chi-square statistic	Discrete random variates
Minimum ESS (*least squares*)	Deterministic data
Minimum average deviation	Deterministic data

Some of the reasons why these five are desirable criteria were described in the last section. It was taken for granted there that, given an appropriate analytical form for the model, finding parameters that would optimize the chosen metric was practicable. Indeed, most of the values for model parameters given in the text are illustrative.

The methodology described here takes a modeling criterion as a kind of meta-parameter. Partial justification for this approach derives from the fact that these five criteria are more interrelated than they might appear. In particular, as first demonstrated by Carl Friedrich Gauss (1777-1855) and Adrien Marie Legendre (1752-1833), there are deep, fundamental connections linking the criteria for stochastic and deterministic modeling. These relationships are due to the randomness of error.

Maximum Likelihood Redux

In the course of any description of modeling and statistics, it is easy to get lost in the almost inevitable deluge of symbology and equations and, in the confusion, lose sight of the original objective. One is dealing, presumably, with real data from the real world. Thus, a model is not just some computational gimmick for reproducing a matrix of numbers. A good model says something genuine about the Universe. Its analytical form is discovered, not invented; its parameters are properties of Nature, not the whim of an analyst.

The ML criterion acknowledges this role, as well as the supremacy of data, in the most direct fashion. Given the form for a stochastic model, ML parameters maximize the probability that the observed data came from a parent population described by the model. Deterministic modeling, on the other hand, attempts to maximize the explanatory power of the model or, occasionally, to minimize the average deviation. It turns out that there is a profound connection between least squares (also, minimum deviation) and maximum likelihood.

We shall start with least squares. This is undoubtedly the most widely used of all modeling criteria. It is usually presented simply as a technique for generating a curve that is closest to all of the data points. Of course, it isn't. The minimum-deviation criterion is the one that yields that particular curve. As we have seen, the least-squares technique actually minimizes the error **variance** and, hence, maximizes the fraction of information that is explained by the model. If it did nothing else, one could hardly complain. However, it does much more.

In real life, errors are almost never simple manifestations of a single mechanism. Nearly always, the total error in an observation is the net effect of several different errors. As a result, the Central Limit Theorem takes over and stipulates that such errors have a strong tendency to be ~Normal(A, B). This theorem was illustrated in Example S3.

In devising a deterministic model, suppose that this theorem and the proliferation of error modes are taken at face value. Then one might ask, "What set of parameter values maximizes the likelihood of the residuals if they are normally distributed? Also, is this vector unique?"

Thus, assume that the residuals for the data points, $\varepsilon_k \equiv \hat{y}_k - y_k$, are ~Normal($A_k$, B_k). In addition, assume that these residuals are all unbiased so that A_k is always zero. Then, the likelihood is given by Equation 19 (cf. Eq. 10 and pg. A-85).

$$\text{Likelihood} \equiv \prod_{k=1}^{N} f\left(\varepsilon_k\right) = \prod_{k=1}^{N} \left[\frac{1}{B_k \sqrt{2\,\pi}} \exp\left(-\frac{1}{2} \left(\frac{\varepsilon_k}{B_k} \right)^2 \right) \right] \qquad 19.$$

Equation 19 can be simplified. The log() function is monotonic, so any set of parameters that maximizes log(likelihood) will necessarily maximize the likelihood. Taking (natural) logs of both sides of Equation 19 produces Equation 20.

$$\text{Log-likelihood} = -\sum_{k=1}^{N} \log\left(B_k \sqrt{2\,\pi} \right) - \frac{1}{2} \sum_{k=1}^{N} \left(\frac{\varepsilon_k}{B_k} \right)^2 \qquad 20.$$

Given a set of measurement errors, the first term in Equation 20 is a constant and has no bearing on the maximization. The likelihood will be a maximum if and only if the expression in Implication 21 is a minimum.

$$\text{Maximum likelihood} \Rightarrow \min\left[\sum_{k=1}^{N} \left(\frac{\varepsilon_k}{B_k} \right)^2 \right] \qquad 21.$$

If the errors for all of the points come from the **same** normal distribution, then all B_k are equal and, after canceling the units, this constant can be factored out giving Implication 22.

$$\text{Maximum likelihood} \Rightarrow \frac{1}{B}\min\left[\sum_{k=1}^{N}\varepsilon_k^2\right] \Rightarrow \min\left[\sum_{k=1}^{N}\left(\hat{y}_k - y_k\right)^2\right] \qquad 22.$$

However, the term on the far right is just ESS! In other words,

> **if all errors are ~Normal(0, B), then the parameters that minimize ESS are the very same parameters that maximize the likelihood of the residuals.**

It turns out that this vector of parameters is almost always unique. The least-squares parameters have many other useful properties as well but these lie beyond our present scope.

While this is a wonderful result, it falls short of perfection for at least two reasons. First, the fact that the parameters are the ML parameters says nothing about the analytical form of the model. If the model itself is inappropriate, its parameter values are academic. Second, there is no guarantee that the errors are independent or normally distributed. These hypotheses must be proven for the least-squares method to yield the ML parameters.

However, the strong assumption that all B_k are equal is unnecessary **provided that their true values are known**. Minimizing Implication 21, instead of Implication 22, is referred to as the *weighted least-squares* technique.[22]

Finally, what about the minimum-deviation criterion? Can that be related to the maximum-likelihood function as well?

Yes, it can. The second most common distribution for experimental errors is the Laplace distribution (see pg. A-63). Substituting the Laplace density for the Normal density, the derivation above produces Implication 23 instead of Implication 21.

$$\text{Maximum likelihood} \Rightarrow \min\left[\sum_{k=1}^{N}\frac{\left|\hat{y}_k - y_k\right|}{B_k}\right] \qquad 23.$$

Implication 23 is equivalent to minimizing the (weighted) average deviation, analogous to weighted least squares. Of course, the same caveats apply here as in the case of normal errors.

Alternate Criteria

The two criteria that do not relate directly to maximum likelihood are the minimum K-S criterion and the minimum chi-square criterion. Nevertheless, the first of these is justified because it is the most common goodness-of-fit test for continuous random variates, as well as one of the most powerful. The chi-square criterion is the discrete counterpart of the K-S criterion and any discussion of discrete variates would be very incomplete without it.

---

[22] Some texts discuss only the latter since setting all B_k equal to one is always a possibility.

From this point on, we shall take it as given that the criteria listed in Table 7 are not only valid and appropriate but, in many respects, optimal. The next task is, therefore, to compute the parameters that realize the selected criterion.

Searching for Optima

As noted earlier, a single computational technique will suffice regardless of the selected criterion. All optimizations of this sort are essentially a search in a multidimensional parameter space for values which optimize the *objective function* specified by this criterion. There are many generic algorithms from which to choose. One of the most robust is the *simplex method* of Nelder and Mead.[23]

The simplex algorithm traverses the parameter space and locates the **local** optimum nearest the starting point. Depending upon the complexity of the parameter space, which increases rapidly with dimensionality, this local optimum may or may not be the **global** optimum sought. It is, as always, up to the analyst to confirm that the computer output is, in fact, the desired solution.

The simplex is a common algorithm and source code is readily available [CAC84, PRE92].

Summary

Three of the five modeling criteria described in the first section have been shown to be closely linked to the concept of maximum likelihood. Consequently, they are eminently suitable as modeling criteria. They are intuitive, valid, and appropriate. Moreover, they almost always provide unique parameter vectors for any model with which they are employed.

Two other criteria were also listed. Both of these are best known as goodness-of-fit criteria for random variates with respect to some putative population density. However, they may also be employed to determine the parameters for this density, not just to test the final result.

The simplex algorithm will prove suitable, in all cases, to compute parameters satisfying the chosen criterion in any modeling task that we shall address. The truth of this assertion has not yet been demonstrated but it will become evident in the sections to follow. Some of the figures shown above provide examples.

To this point, we have seen, or at least stated, how models may be specified, optimized, and computed. The stage is now set to evaluate the results of any such endeavor. We shall discuss the quality of models in general and of their parameters in particular. Is the model any good? Are the parameters credible? Does the combination of model and parameters make sense in the light of what we know about this experiment and about the Universe in general? In other words, are we doing real science/engineering/analysis or are we just playing with data?

---

[23] not to be confused with the linear programming technique of the same name

IS THE MODEL ANY GOOD?

The "goodness" of a model depends not on how well it might serve our purposes but on the degree to which it tells the truth. Beauty, in this case, is not entirely in the eye of the beholder. If a model is based upon observed data, especially physical data about the real world, then the model must be equally real. Several criteria have been established, in earlier sections, that measure the validity (i.e., the reality) of a model. In this section, we shall focus mainly on the statistical dimension of this reality; the physical dimension is beyond our scope.

Deterministic data and stochastic data will be treated separately because the former case is very easy but the latter is not. However, as we have seen, it is the modeling of error as a random variable that lends credence to the aforementioned deterministic criteria. Thus, the validity of stochastic models is fundamental to the validity of modeling in general.

In this section, we encounter a new technique—the *bootstrap*. The bootstrap may be *parametric* or *nonparametric*. In the discussions to follow, both forms will be employed. The utility of these techniques should be self-evident.

The Bootstrap

Since there is ample literature on bootstrap methodology [EFR93], the descriptions here will be somewhat abbreviated. While the name is fairly new, some of what is now termed "bootstrap" methodology is not. In particular, the parametric bootstrap includes much that was formerly considered simply a kind of Monte Carlo simulation. However, the nonparametric bootstrap, outlined in the next section, is quite recent.

The bootstrap technique was designed, primarily, to establish *confidence intervals* for a given statistic. A confidence interval is a contiguous range of values within which the "true" value of the statistic will be found with some predetermined probability. Thus, the two-sided, 95-percent confidence interval for a given K-S statistic is that range of values within which the true K-S statistic will be found in 95 percent of all samples drawn from the given population, being found 2.5 percent of the time in each tail outside the interval. Other things being equal, the width of this interval decreases with increasing sample size.

Sometimes, a confidence interval may be computed from theory alone. For instance, means of large, random samples tend to be unbiased and normally distributed. Therefore, the 95-percent confidence interval for any such mean is just $\mu \pm 1.96$ SE, where μ is the observed mean and SE is the *standard error* of the mean, given by Equation 24.

$$SE = \frac{\sigma}{\sqrt{N}} \qquad 24.$$

where σ is the population standard deviation for the data and N is the sample size.

As an example, consider again the data exhibited in Figure 7. Theory says that these 1,000 means should, themselves, have an (unbiased) mean of 18 and a standard deviation (SE) of 0.30. Hence, 95 percent of them should lie in the interval $18 \pm 1.96*0.30 = [17.41, 18.59]$, with the remaining five percent divided equally between the two tails. In fact, if you sort the file *Examples:Rolls3avg.in*, you will find that these 1,000 means exhibit a 95-percent central

confidence interval of [17.40, 18.55]. The latter interval is, of course, just a sample of size one of all similar confidence intervals for similar data. As such, it is expected to deviate a little bit from theory. If you sense a little recursion going on here, you are quite right.

What we have just done, without saying so, was a form of parametric bootstrap. The file of 1,000 means was obtained by simulating the original experiment 1,000 times. Each trial was terminated when an event with a probability of **1/6** was realized **3** times. The simulation used these two **parameters** to synthesize 1,000 *bootstrap samples*, each sample containing 1,000 trials, with only the sample mean recorded. The explicit use of input parameters makes this a parametric bootstrap.

Figures 7 and 8, and the accompanying discussion, illustrate not only the Central Limit Theorem but also the validity of the bootstrap technique. This technique has received a great deal of scrutiny over the past decade, with very positive results. We shall put it to good use. Almost all of our remaining discussion will derive from the now well-established fact that

bootstrap samples are valid samples.

Whether they are created from theory (parametric bootstrap) or experiment (nonparametric bootstrap), bootstrap samples constitute **additional** data. These data may be characterized by one or more statistics and the ensuing distribution of these statistics utilized to estimate any desired confidence intervals. By comparing optimum values, for any model, to the confidence intervals for these values, the probability of the former may be estimated. In this section, this probability will be related to goodness-of-fit; in the following section, it will be employed to estimate parameter uncertainty.

Stochastic Models

Table 7 listed three criteria for optimizing a stochastic model. Unlike their deterministic counterparts, none of these criteria are self-explanatory. In general, it is not easy to tell, from the value of the given statistic, whether or not that value is probable under the hypothesis that the optimum model is valid.

For example, the K-S statistic for the best-fit Gumbel distribution to the Batting-Average data is 0.057892. This number represents the maximum (unsigned) discrepancy between the theoretical and empirical CDFs. So much is very clear. What is not at all clear, however, is whether or not **a random sample (N = 97) from a genuine population described by the same distribution** would be likely to exhibit a K-S statistic having the same value. It could be that the value 0.057892 is too large to be credible.

How could you tell? The traditional approach would be to look up a table of K-S critical points in a reference and compare the observed value to some table entry. The problem with this approach is that nearly all such tables list asymptotic results, i.e., with N = Infinity. Even those that have entries for specific sample sizes are not always appropriate in every case. There might be hidden dependencies on model form or parameters that could invalidate the recorded critical points. Ideally, what one would really like is a made-to-order table, conditioned on the sample size, the model form, and the model parameters. Before the advent of fast digital computers, such a table was a rare luxury. Today, it is fairly easy.

To construct such a customized table, we start with the putative parent population. In this case, we start with the Gumbel distribution given in Equation 25.

$$\text{PDF} = \frac{1}{0.01984}\exp\left(\frac{0.3555 - y}{0.01984}\right)\exp\left(-\exp\left(\frac{0.3555 - y}{0.01984}\right)\right) \qquad 25.$$

From this distribution, which purports to be a good description of our data, draw a random sample (N = 97) a very large number of times (at least 1,000 times). With each of these samples, **carry out the same operations that were performed on the original data**. Here, we assume that each sample is ~Gumbel(A, B) and we optimize the model, for each sample, using our original criterion. When all of this is done, we have a vector of 1,000 **bootstrapped** K-S statistics, all of which a) genuinely reflect our candidate population, by construction, and b) were obtained exactly as was the original K-S statistic.

There are now several ways to proceed but, here, we shall adopt the simplest approach. First, sort the 1,000 bootstrapped statistics in ascending order. The result will look much like the histogram in Figure 12 (where Y = K-S statistic).

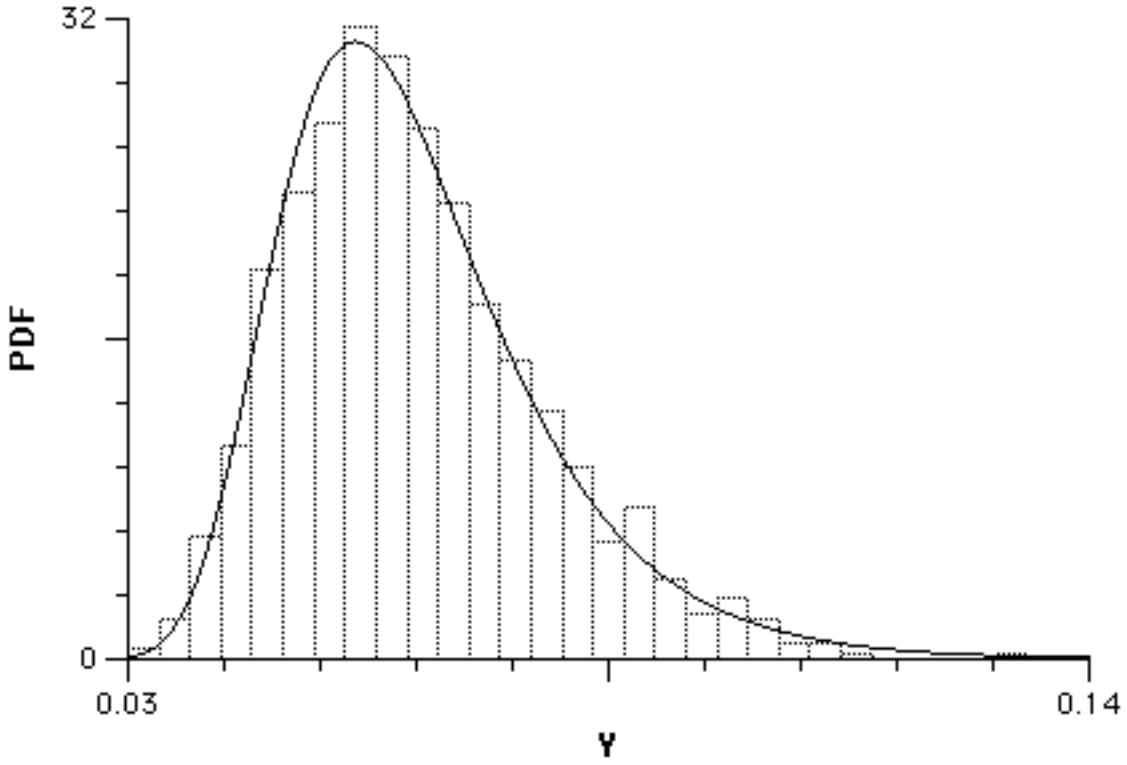

Figure 12. Distribution of K-S Statistic from Equation 25

The bootstrapped K-S statistic is, itself, a random variable and we could try to model it. Since each of these continuous variates represents the maximum of 2 x 97 comparisons,[24] they might even be ~Gumbel(A, B)![25] However, we shall instead use just the raw values. It turns out that our observed value of K-S is found in the 41st *percentile* of this specific K-S distribution. In other words, if you draw a sample (N = 97) from the distribution given by Equation 25, model it as ~Gumbel(A, B), and recompute the optimum A and B using the ML criterion then, 59 percent of the time, the K-S statistic you will get will be larger **(worse)** than the one we obtained with the Batting-Average data.

It is customary to require the 95th percentile before rejecting an hypothesis which means that we **cannot** reject the null hypothesis that our data are modeled by Eq. 25. While not rejecting an hypothesis is not quite the same thing as accepting it, in practice, this is the usual interpretation. Therefore, we conclude that

according to the K-S test, Eq. 25 is an ACCEPTABLE description of our data.

We could go through the same procedure with the 1,000 Log-likelihood statistics that resulted from this bootstrap process, assessing the observed value in the same fashion. Then, we would have two measures of acceptability. The prudent course would be to require both of them to be acceptable before declaring that the model is good. In this example, the observed value falls in the 47th percentile of its (bootstrapped) distribution so it, too, is acceptable.

With discrete variates, the usual goodness-of-fit statistic is Chi-square. The same procedure would apply there as well. In fact, the bootstrap technique is a very general method for assessing goodness-of-fit, no matter what statistic/criterion is used.

There is also a simple rule-of-thumb when using the Chi-square statistic as described. Its expected value is equal to the number of degrees of freedom.[26] Therefore, if the Chi-square value found is less than the range of y-values, it will always be acceptable.

Deterministic Models

Since the bootstrap technique is generally valid, it could be used for assessing goodness-of-fit with deterministic models as well. However, R-squared, the usual statistic in such cases, is self-explanatory. It is equal to the fraction of the variance explained by the model. There is no need to ask about the probability of such a value with a given dataset/model combination.

The minimum-deviation statistic is not quite as transparent. Still, this statistic has the same units as the data and it is usually easy to tell if a given minimum deviation is small enough for the intended purposes.

---

[24] Each end of each "step" in the empirical CDF must be examined.

[25] Indeed they are, Gumbel(0.05604, 0.01192) [Fig. 12, solid line]. In fact, this model is an even better fit than it was with the Batting-Average data. See also, *Examples:KStest.in*.

[26] see pg. A-42

All of the above is true whether the deterministic modeling (*regression*) is weighted or unweighted. With weighted regressions, the weighted residuals take the place of the usual residuals since the former are now normalized with respect to their own standard errors. However, the goodness-of-fit metrics, and their meaning, remain the same.

Summary

This section has outlined a general technique for establishing the validity of a mathematical model. If a model is deemed acceptable, it signifies that the equation (or density function) may substitute for the original data. Moreover, the model may now be queried, perhaps in ways that would be impossible with the actual dataset. Deterministic models may likewise be used to interpolate or extrapolate from the observed values. In short, if a model is acceptable, it is equivalent to any and all past (or future) observations.

With stochastic models, acceptability is assessed by comparing one or more goodness-of-fit statistics (random variables) to their respective distributions, asking whether the value observed is too improbable to credit. With deterministic models, the goodness-of-fit statistics are sufficiently perspicuous that we need not go to so much trouble.

Once again, it must be emphasized that none of this is meant as some kind of mathematical gimmick. The number of potential models is infinite and, with enough fortitude, any dataset could be adequately fit to one or more of them. The real goal is to find the model that Nature uses by examining a sample of observations. The techniques described herein are immensely powerful but they are blind. The necessary vision must be supplied by the analyst.

HOW PRECISE ARE THE PARAMETERS?

Our goal has been to show how to obtain a validated mathematical model for any dataset. This we have done. Yet the story is not quite complete. Certainly, an acceptable model of the proper form, with optimum, ML parameters cannot be significantly improved but there remain two loose ends.

One loose end is that the procedure described above for determining ML parameters is just a search of the parameter space for the best parameter vector. In general, there is no guarantee either that there is only one best solution or, if there is, that you will find it. Depending upon where you start looking, you might converge on a local optimum instead of the desired global optimum. Also, the latter might be degenerate, i.e., there could be more than one of them.

For instance, the ML fit of a sine wave to the Daytime data (Table 1) is shown in Figure 13 (cf. Figure 1). This fit (Eq. 26) has R-squared = 0.99908 so this model is good to one part in a thousand but simple, high-school trigonometry reveals that the phase term, –1.391, may be incremented by any integer multiple of 2π (360 degrees) without changing the daytime value, y. Therefore, the number of ML solutions is actually infinite.

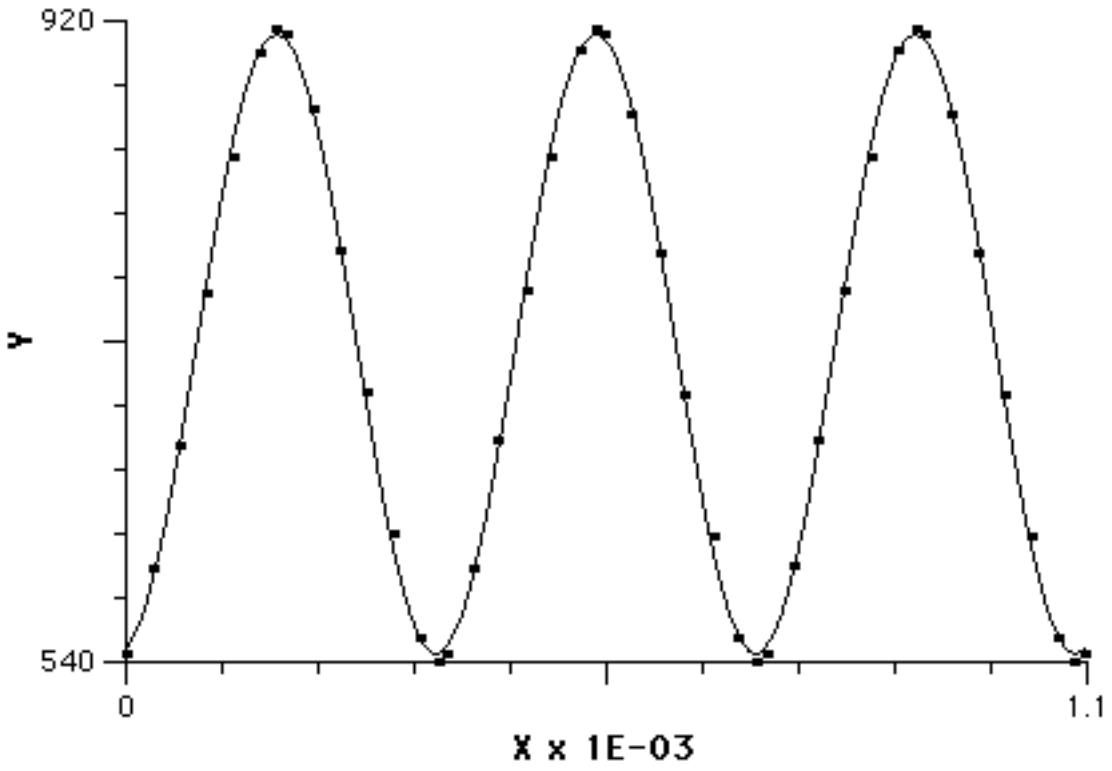

Figure 13. Least-squares Sine Fit to Daytime Data

$$y = 183.3 \sin((2\pi)\, 0.002736\, x - 1.391) + 728.4 \qquad 26.$$

The other loose end stems from the definition of maximum likelihood. ML parameters are, by definition, more probable than any others—but how probable is that? Suppose you propose to flip a coin 100 times and bet on the outcome. The wise bet would be 50 Heads and 50 Tails because that is the ML result; every other result is less likely. Suppose, however, that some bystanders are watching all of this and decide to bet, amongst themselves, whether or not you will win your bet. The wise choice in this case would be to bet that you would lose. Even though, with a fair coin, a 50:50 outcome is the most probable, its probability is still less than eight percent. There are just too many other possibilities for any one of them to be at all likely. So your chances of losing are, at best, more than 92 percent.

The probability that the ML parameters in a model are equal to the true parameters is even worse. Unless theory insists that a given parameter must be a rational number, then you can be absolutely certain that the optimum value you found for it will always be incorrect. The reason, noted earlier, is that selecting a specific number from the real domain has a probability of zero. You can never do it exactly.[27]

But you can get as close as you like. All it takes is a lot of effort and a **very** large sample.

Sampling is the key. The model for the radius of the orbit of Mars (Table 5, col. 3) as a function of true anomaly has two ML parameters: A = 1.51043 AU and B = 0.0926388. The Keplerian model is an approximation, albeit a very good one, to the true solution. Nevertheless, even if we grant that the elliptical approximation is good enough, these two parameters are not perfect. Were new data found for exactly the same time frame, these parameters would be different **even though the orbit of Mars was constant**. Every dataset is a sample, not the entire population. Therefore, there is always some sampling error.

Nothing can be done to compensate for degenerate global optima except to examine the ML parameters for reasonableness. They must make sense! They must have the correct units and be consistent with what is known about the real world. This is essential even if the solution is not degenerate. Likewise, parameter uncertainties due to sampling error, or to deficiencies in the model, cannot be removed once the ML result has been obtained. However, they can be estimated.

The purpose of this section is to describe a procedure for estimating parameter uncertainty. In addition, parameter correlation is also discussed. The latter is a subtle kind of uncertainty which recognizes that parameters are not necessarily independent. We shall take it as given that the model is acceptable since a model that is not acceptable is of little interest and its parameters even less so.

To quantify uncertainty, we shall utilize confidence intervals. Every optimum parameter can be assigned a confidence interval within which the true value of the parameter will be found with some specified probability. The wider this confidence interval, the more uncertain the parameter value.

---

[27] assuming that you cannot go out and examine every member of the parent population

As we have already seen, confidence intervals may be estimated *via* the bootstrap. For this application, however, the parametric bootstrap is not suitable. We are not asking now about a theoretical population from which we can select random variates. Rather, we are asking about one particular dataset, the one that gave rise to the model. Hence, we need a **nonparametric** bootstrap that somehow uses this dataset to generate further bootstrap samples. The requisite methodology is outlined below. As we have done throughout this tutorial, we concentrate on examples instead of theory, first with stochastic models, then with deterministic models.

The Nonparametric Bootstrap

A nonparametric bootstrap sample is created by **resampling the original dataset, with replacement**. Imagine that the original data are in a paper bag. Close your eyes and select one point/variate from the bag, record its value, then put it back. Repeat this selection until you have a sample of the same size as the original. Such a sample is nonparametric; it comes from the actual data, not from a parametric model of that data. Reusing the data in this way might seem like cheating[28] but it is quite valid, as long as proper corrections are applied.

In the previous section, when we discussed the parametric bootstrap, a simple approach was used to analyze the bootstrap distributions. For instance, consider the 1,000 K-S statistics in Figure 12. These 1,000 values constitute an empirical CDF for any K-S statistic computed for a sample (N = 97) that is ~Gumbel(0.3555, 0.01984). To determine a central, 95-percent confidence interval for K-S, using the **percentile** method, we throw away the 25 largest and 25 smallest of these 1,000 K-S values. The remaining 0.95*1,000 = 950 values comprise the desired confidence interval, conditioned on this model and the given sample size. Its range is [0.0401747, 0.0974331]. Its interpretation is that 95 percent of the K-S statistics you would obtain with a genuine sample (N = 97) from this Gumbel distribution would fall in this range. Furthermore, 2.5 percent would fall in the range [0, 0.0401747] and the last 2.5 percent in the range [0.0974331, 1].

Were the model or sample size different, the 95-percent confidence interval would be different. The confidence interval would also grow or shrink with the need for increased or decreased confidence, respectively. Thus, the 90-percent confidence interval in this example is [0.0430050, 0.0901146] and the 99-percent confidence interval is [0.0364837, 0.107393].

With the nonparametric bootstrap, it is better **not** to use such a simple approach to find a confidence interval. When the original data are resampled, the resulting bootstrap distribution is biased and skewed. The corrections referred to above compensate for these defects. Instead of using the 25th and 975th elements of the sorted CDF to demarcate the tails of the bootstrap distribution, the *BCa technique* specifies alternate indices.[29] Once these have been identified, however, the interpretation of the confidence interval remains the same.

We shall now illustrate the use of these nonparametric confidence intervals to determine the precision (confidence) that we may ascribe to the ML parameters of any model.

---

[28] or, at least, "pulling oneself up by one's bootstraps"

[29] Computing the latter is well understood but rather complicated [see *Technical Details*].

Stochastic Models

Repeat the following procedure 50 times:

Step 1
Let X and Y be random Cartesian coordinates ~Normal(0, 3).
Step 2
Select X and Y, then draw the radius, R, from the origin to the point (X, Y).
Step 3
Record the length of R.

At the end of this experiment, we have a sample of 50 values for R.[30] Although we shall not prove it, theory says that R is ~Rayleigh(3) which is the same as ~Chi(3, 2).[31] However, let us pretend that we know that the distribution should be Rayleigh but have forgotten how to compute the theoretical scale parameter, A. Instead, we shall just determine the ML parameter for a Rayleigh(A) model (Equation 27), based upon this small sample.

$$\text{PDF} = \frac{y}{A^2} \exp\left(-\frac{1}{2}\left(\frac{y}{A}\right)^2\right) \qquad 27.$$

The result (see Figure 14, with Y representing R) is quite acceptable, but with A = 2.847 instead of 3. What happened? Did we make a mistake? Is the theory wrong? Was our pseudo-random number generator faulty?

The answer is D, none of the above. Our 50 variates are just one sample out of an infinite number of similar samples that this experiment could have produced. It is extremely unlikely that the ML parameter of the ~Rayleigh model for any such sample would equal the theoretical prediction to four significant figures. Had the sample size been 1,000,000,000, then perhaps four significant figures might have been anticipated, but not with N = 50.

We can be certain of this because, using 2,000 nonparametric bootstrap samples, we find that the estimated 90-percent central confidence interval for A is [2.648, 3.062]. Since even this somewhat conservative (narrow) confidence interval includes the value 3,

we cannot reject the null hypothesis that A = 3.

Given the manner in which the 50 radii were created, they necessarily constitute a **genuine** sample (N = 50), ~Rayleigh(3). The confidence interval given above, therefore, reflects the intrinsic variability of this parameter. Were another sample of 50 radii constructed, its ML parameter would be different from that found for the first sample but the confidence interval would be much the same. The variability of confidence intervals decreases as the number of bootstrap samples increases. A set of 1,000 bootstrap samples is the minimum for experiments such as these.

---

30 see file *Examples:Radii50.in*

31 see pg. A-21

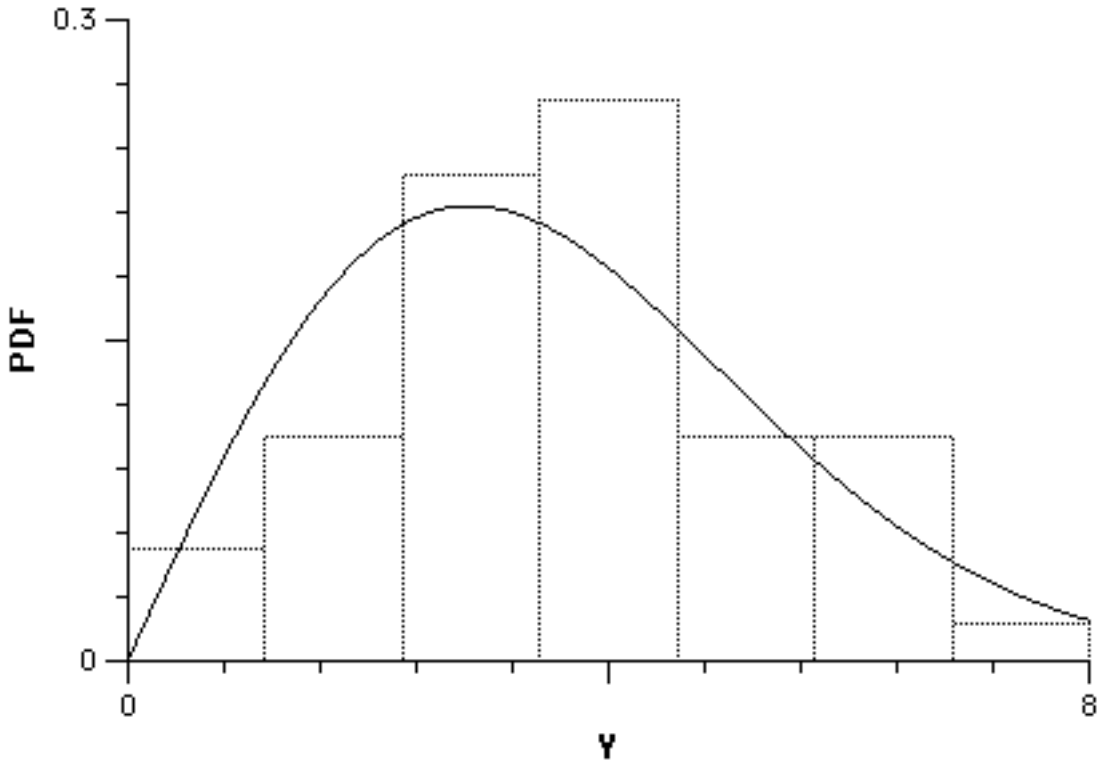

Figure 14. ML Fit of 50 Radii to a Chi Distribution

If we repeat this experiment with 5,000 nonparametric bootstrap samples instead of 2,000, in order to achieve greater accuracy, we obtain the following results for parameter A:

90% --> [2.649, 3.072]
95% --> [2.597, 3.138]
99% --> [2.507, 3.264]

At the same time, the corresponding parametric confidence intervals, determined from the model and the percentile method, instead of by resampling the data and the BCa method, are as follows:

90% --> [2.527, 3.186]
95% --> [2.460, 3.246]
99% --> [2.327, 3.355]

Finally, it is interesting to compare our sample of 50 radii to its theoretical distribution. If we insist that A = 3, we get the theoretical PDF shown in Figure 15 (with Y representing R) along with the empirical histogram representing the data. Using the ML and K-S criteria established in the previous section, plus 2,000 parametric bootstrap samples, it turns out that this fit is likewise acceptable, with likelihood and K-S values falling in the 9th and 35th percentiles, respectively.

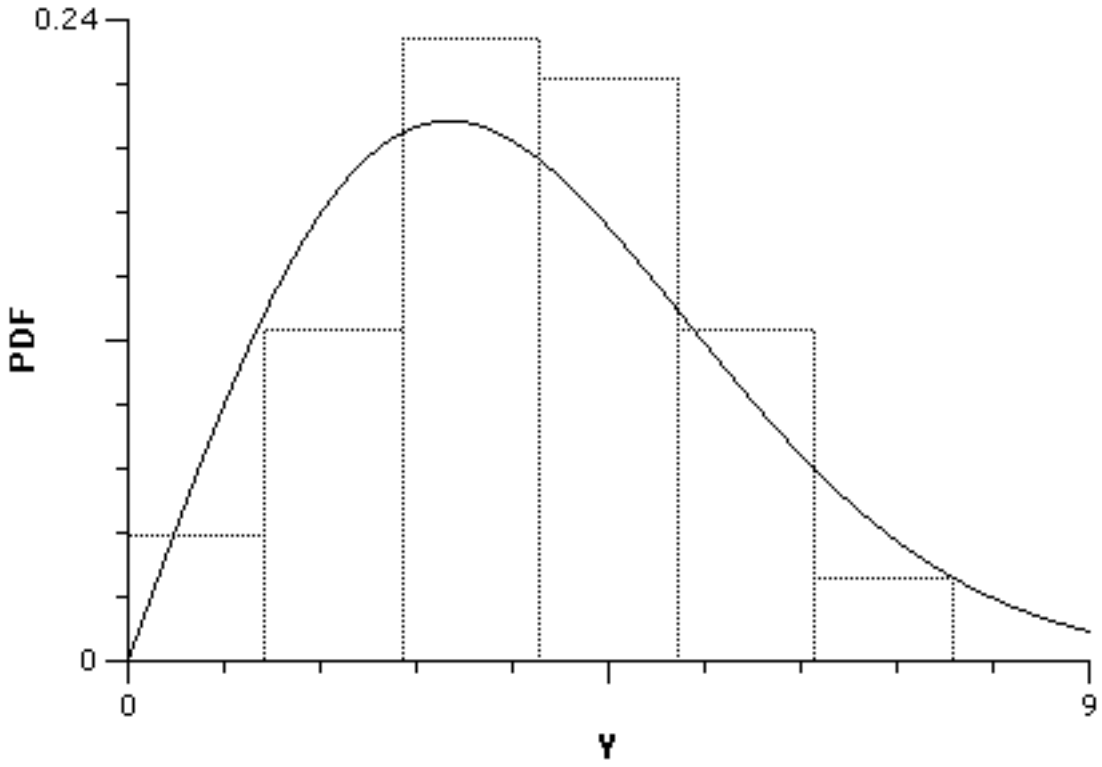

Figure 15. Comparison of 50 Radii to their Theoretical Rayleigh Distribution

Deterministic Models

Consider the data and model shown in Figure 13 and Equation 26. These daytime values were modeled with a sine function having four parameters. The value of R-squared, 0.99908, is excellent and we would have no qualms about accepting such a fit. Still, by using the least-squares technique, we implicitly assumed that the residuals were ~Normal(0, B). We did not prove it. Now that we know how to assess a distribution of random errors, let us do so.

Table 8 shows again the daytime data of Table 1, along with the daytime values predicted by our supposedly optimum model. A brief glance at either this table or at Figure 13 discloses that the fit is quite good. The residuals are listed in column 4. The worst one, shown in bold, is only –7.2 minutes. The parameters given in Equation 26 are obviously very good. However, they are the ML parameters if and only if we can prove that the vector of residuals shown here is ~Normal(0, B). Otherwise, as good as the fit is, there are better parameters yet to be found, perhaps even a better model.

We shall not repeat earlier discussion. As Sherlock Holmes would say, "You know my methods." We find that the residuals are best described as ~Normal(0, 4.280). In this case, the mean was held constant at zero, not estimated from the data. The likelihood and the K-S statistics fall in the 49th and 82nd percentile, respectively, based upon 1,000 (parametric) bootstrap samples. The unbiased, Gaussian nature of these residuals is, therefore, accepted.

Table 8. Daytime Data, Observed and Modeled

Observed		Modeled	
Daytime (min.)	Day	Daytime (est.)	Residual
545	1	548.7	3.7
595	32	591.8	–3.2
669	60	664.0	–5.0
758	91	760.1	2.1
839	121	845.0	6.0
901	152	900.7	–0.3
915	172	911.7	–3.3
912	182	909.2	–2.8
867	213	868.6	1.6
784	244	789.2	5.2
700	274	696.0	–4.0
616	305	**608.8**	**–7.2**
555	335	555.9	0.9
540	356	545.1	5.1
544	366	548.4	4.4
595	397	590.8	–4.2
671	426	665.5	–5.5
760	457	761.7	1.7
840	487	846.3	6.3
902	518	901.3	–0.7
915	538	911.7	–3.3
912	548	908.9	–3.1
865	579	867.6	2.6
782	610	787.7	5.7
698	640	694.4	–3.6
614	671	607.6	–6.4
554	701	555.3	1.3
540	721	545.1	5.1
545	732	548.7	3.7
597	763	591.9	–5.1
671	791	664.0	–7.0
760	822	760.1	0.1
839	852	845.1	6.1
902	883	900.7	–1.3
915	903	911.7	–3.3
912	913	909.2	–2.8
865	944	868.6	3.6
783	975	789.1	6.1
699	1005	695.9	–3.1
615	1036	608.8	–6.2
554	1066	555.8	1.8
540	1086	545.1	5.1
545	1097	548.4	3.4

In saying this, we imply only that the residuals, collectively, could easily have come from a zero-mean Normal distribution. "Collectively" signifies that we do not consider the **order** of the residuals with respect to each other. If you look carefully at Table 8, column 4, you will notice a suspicious alternation of positive and negative residuals. Such a periodicity would **not** be observed in a random sample from any Gaussian distribution. It is, rather, a **systematic error** in our model. It appears that we have not, after all, completely characterized the motions of the Earth and Sun with this simple equation. Kepler could have told us that!

We forego perfection for the moment and take up the matter of uncertainty in the least-squares parameters. To do this, we shall again require a large number of nonparametric bootstrap samples. There are two common methods for obtaining such samples. We could simply resample the data points as we did with the 50 radii of the previous example. However, it is more common to use all of the data points and resample the residuals instead. If the latter are independent (hence, uncorrelated), then any residual might have been found associated with any data point. Here, there is some small amount of correlation but, for the sake of this illustration, we shall ignore it and proceed in the usual fashion, using 1,000 bootstrap samples.

Doing so, we obtain simultaneous confidence intervals for all of the parameters, as follows:

A 90% --> [182.247, 184.318] (cf. Eq. 26, page 37)
95% --> [181.944, 184.594]
99% --> [181.354, 184.991]

B 90% --> [0.00273247, 0.00273958]
95% --> [0.00273152, 0.00274036]
99% --> [0.00272935, 0.00274223]

C 90% --> [–1.40549, –1.37757]
95% --> [–1.40857, –1.37274]
99% --> [–1.41479, –1.36706]

D 90% --> [727.679, 729.264]
95% --> [727.465, 729.569]
99% --> [727.030, 729.939]

Note that the widths of these intervals, relative to parameter magnitude, are much narrower (i.e., better) than those seen in the example of the 50 random radii in spite of the fact that the latter were a genuine sample from a known parent population while the sine model here is known to be deficient.

It is difficult to predict the variability of parameters. It was not obvious, for instance, that the uncertainty in the period of these data would be less than the uncertainty in the amplitude. Unless confidence intervals can be computed from theory, only a large bootstrap sample will yield this sort of information.

The interpretation of the confidence intervals is the same with these deterministic data as it was with the stochastic data in the previous example. Also, the accuracy of the intervals will improve, *ceteris paribus*, with bootstrap sample size.

Parameter Correlation

There is one, final comment to make concerning confidence intervals and that concerns their appearance. A confidence interval is just a range of numbers and it is very easy to fall into the trap of believing that all of the numbers in that range are equally probable or, if there is more than one such range for a given model, that any parameter may have any value in the stated range regardless of the values of the other parameters. Neither of these beliefs is valid.

As an illustration, let us examine the nonparametric bootstrap results for the daytime data more closely. Given 1,000 bootstrap samples, we may compute the mean of all the bootstrap parameters and, also, the covariance matrix.

The means are as follows:

A = 183.383
B = 0.00273601
C = –1.39028
D = 728.435

Note, first of all, that these means are not quite equal to the ML values of the parameters. Bootstrap distributions are often unsymmetrical, with the mode and mean unequal.

Here is the upper triangle[32] of the covariance matrix, with indices from A to D:

6.82581e–01	2.12136e–08	1.06586e–03	5.41600e–02
	8.03571e–12	–2.29840e–08	4.54181e–07
		1.19938e–04	–6.43826e–04
			4.14830e–01

The ijth entry in this matrix is given by Equation 28:

$$\mathrm{Cov}\left(P_{ij}\right) \equiv \sigma_{ij}^2 = \left\langle \left(P_i - \left\langle P_i \right\rangle\right)\left(P_j - \left\langle P_j \right\rangle\right) \right\rangle \qquad 28.$$

where P is a parameter.

When i = j, this is just the variance of parameter[j]; when i ≠ j, it is the *covariance*. As we have seen before, the variance measures the intrinsic variability of P. The covariance measures the variability of one parameter with respect to another. Looking at this matrix, we see that parameter B, the period, has the smallest variance. No matter what the other parameters might have been, for any of the 1,000 bootstrap samples, the period stayed almost constant.

The sign of the covariance is an indication of the sense of correlation. Covariances that are positive indicate that the two parameters in question increased and decreased together, on average. Negative covariances mean that, on average, one parameter increased when the other decreased. Here, for example, the amplitude of the sine wave, A, is positively correlated with

[32] All covariance matrices are symmetric about the main diagonal.

all of the remaining parameters. However, the phase, C, is negatively correlated with D. To get an idea of how strong these relationships are, it is, perhaps, best simply to compare the covariances for any parameter to its own variance. Any statistics textbook will discuss these matters in great detail.

In any case, the implications with regard to confidence intervals is clear:

All combinations of parameters are not equally probable.

Summary

Our goal was to establish a quantitative measure of the variability of model parameters. We have succeeded by taking advantage of the power of the nonparametric bootstrap. The measure of variability obtained is a confidence interval which denotes the numerical range wherein the true value of the parameter is likely to be found with some specified probability. By using information from the covariance matrix, we gain additional insight into the manner in which parameters change with respect to each other when the dataset changes.

99.9999999999% PERFECT!

The Rydberg constant describes the colors of atoms. For a hypothetical atom of infinite mass, its limiting value, R_∞ = 109,737.3156864 cm^{-1}, has an accuracy of more than twelve significant figures [UDE97]. Think about that for a moment.

That's like measuring the average distance between Earth and the Moon with an error less than the width of the period at the end of this sentence. Clearly, this is far beyond our current technology. How, then, can we measure R_∞ with such perfection?

Kepler, and his eponymous laws of planetary motion are partly responsible. Using the first and third of these, it is possible to derive the classical expression for the energy of an electron in an atom. If this expression is combined with the laws of Quantum Mechanics (QM) and its insistence upon the importance of integers, the result is that an atom cannot have just any color but only a limited (yet still infinite!) set of colors.

Evidence for this conclusion can be illustrated using Hydrogen, the simplest atom, with the simplest set of colors. The primary colors of Hydrogen are listed in Table 9.[33]

Table 9. Observed Transitions/Colors for Hydrogen ($_1H^1$)

		Wavelength (in Angstroms)					
	N_l	1	2	3	4	5	6
	Series	Lyman	Balmer	Paschen	Brackett	Pfund	Humphries
N_u		*Ultraviolet*	*Visible*		*Infra-red*		
2		1215.67	6562.72	18751.0	40511.6	74578.0	123685.
3		1025.72	4861.33	12818.1	26251.5	46525.1	
4		972.54	4340.47	10938.1	21655.3		
5		949.74	4101.74	10049.4			
6		937.80	3970.07	9546.0			
7		930.75	3889.05				
8		926.23	3835.38				
9		923.15	3797.90				

The Rydberg constant describes all of these colors *via* the model given in Equation 29.

$$\text{Transition energy} \sim R_H \left(\frac{1}{N_l^2} - \frac{1}{N_u^2} \right) \qquad 29.$$

where N_l and N_u are positive integers and where ~, in this case, means "is proportional to."

According to the laws of QM, the color, i.e., the wavelength (λ) of the light emitted during any transition is inversely proportional to its energy. The final model, based on these laws, is given by Equation 30. Fitting the data in Table 9 to Equation 30, the (unweighted) least-

[33] see file *Examples:Rydberg.in* [WEA75]

squares value for R_H is 109,708 cm^{-1}.[34] The resulting graph is shown in Figure 16, with Y representing wavelength. To six significant figures, R-squared = 1.00000. When experiments along these lines are done very carefully, far more than six significant figures can be achieved. However this is still not enough to get the reported precision of R_∞.

$$\lambda = \frac{x}{R_H} = \frac{1}{R_H} \frac{N_u^2 N_l^2}{N_u^2 - N_l^2} \qquad 30.$$

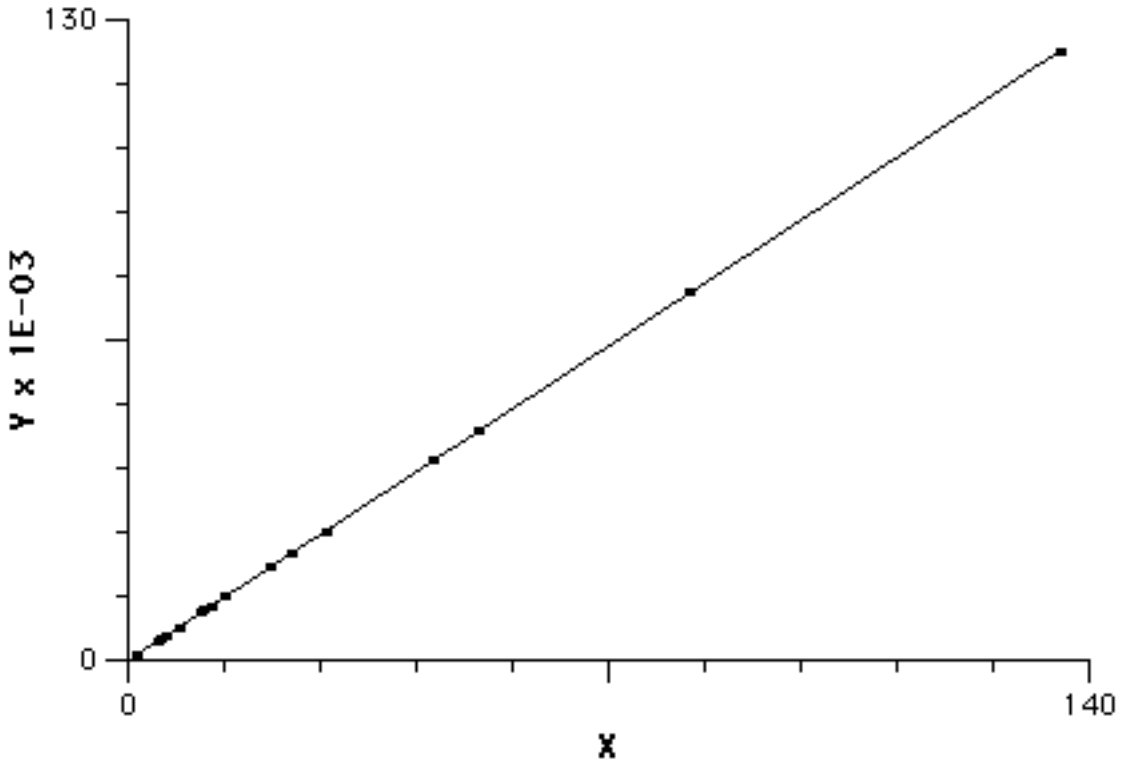

Figure 16. Rydberg Data for Hydrogen

The current accuracy of the Rydberg constant derives not from any single experiment or insight into the laws of Nature but from the unavoidable requirement that the Universe be consistent. This consistency must be mirrored in the physical sciences, specifically in the many scientific models that have been developed over the last four centuries. When all of these models are considered simultaneously, they must agree.

This requirement is manifested in relationships such as Equation 31.

[34] One centimeter = 10^8 Angstroms.

$$R_\infty = \frac{2\,\pi^2\,m\,e^4}{c\,h^3} \qquad 31.$$

where m and e are the mass and charge of the electron, respectively, c is the speed of light, and h is Planck's constant, the proportionality constant connecting energy to wavelength.

Equation 30 relates a mass, a charge, the speed of light, and several other constants, to the observable colors of atoms. One can perform a variety of experiments to determine this mass, charge, or speed. In every case, there are still more factors, determinable in yet different ways. Each of these experiments depends upon a model and all of these models interlock. They are like words in a crossword puzzle. The total solution will break down if even one of them is incorrect.

The physical sciences, which seek this total solution, are not incorrect; they are mature and reliable. One need only turn on a television set or view a CAT scan to appreciate that the laws of the Universe are very well understood. Except at the frontiers of knowledge, the puzzle all holds together. It does not break down.

It is, in part, this internal consistency which affords the degree of accuracy that is seen in R_∞. Whenever any physical experiment is performed, it yields a quantitative answer with some precision. If sufficient attention has been paid to eliminating sources of error, the experiment will have high precision. When many such experiments are incorporated into a single analysis, it is possible to compute a vector of means for any quantities, such as those in Equation 30, that their answers might have in common. As we have seen, in Equation 24 for instance, these means will be more precise than any single experimental result. Simply put, the information from many experiments necessarily exceeds the information provided by just one of them.

In our efforts at modeling, we have sought to extract information from data. The examples described herein illustrate that contemporary methodology is very effective in this task. When the same methodology is applied to a collection of experiments, it is possible to achieve enough synergy to eliminate almost all of the residual error. It is possible to obtain 99.9999999999 percent accuracy.

Twelve significant figures do not come about by accident. The techniques described in this tutorial are just the latest in a heritage that goes all the way back to Kepler and Galileo. That heritage is one which affirms the primacy of data as well as the power of mathematics and statistics. These affirmations are implemented in techniques that are based upon the concept of maximum likelihood, techniques that literally construct the best possible match between theory and experiment, between model and data.

In computing the optimum result, these same techniques explicitly acknowledge the role of measurement error and sampling error. Even the best measurements are not perfect. Even the best experiments must examine less than the entire Universe.

Error implies that the models we construct, and their parameters, must always be somewhat uncertain, however much we might like to believe otherwise. No model is complete until those uncertainties have been rigorously quantified. A good model knows its limitations…

…as does a good tutorial.

We began with data, but where do we end? Perhaps T. S. Eliot had the right idea:

> "We shall not cease from exploration
> And the end of all our exploring
> Will be to arrive where we started
> And know the place for the first time."

And be ready to start again on a new adventure, older and wiser, and better prepared—

to experiment,

to analyze,

to play the game one more time.