Chapter 1 INTRODUCTION

1.1 The Experimental Method

We can consider the experimental method as consisting of four distinct phases:

1) Design
2) Execution
3) Data Analysis
4) Interpretation

The design phase is partially problem dependent. For example, we require answers to questions regarding the choice of equipment and the physical layout of the experimental setup. However, there are also questions of a more generic nature: for example, how many data points are needed and what are the accuracy requirements of the data (i.e., how accurately must we measure or compute or otherwise obtain the data)?

The execution phase is completely problem dependent. The performance of an experiment is usually a physical process although we often see computer experiments in which data is "obtained" as the result of computer calculations or simulations. The purpose of the execution phase is to run the experiment and obtain data.

The data analysis phase is typically independent of the details of the physical problem. Once data has been acquired and a mathematical model has been proposed, the actual analysis is no longer concerned with what the data represents. For example, assume that we have obtained values of a dependent variable $\boldsymbol{Y}$ as a function of time $\boldsymbol{t}$. We propose a mathematical model that relates $\boldsymbol{Y}$ to $\boldsymbol{t}$ and the ensuing analysis considers the values of $\boldsymbol{Y}$ and $\boldsymbol{t}$ as just a collection of numbers.

The interpretation phase is really a function of what one hopes to accomplish. There are a number of reasons why the experiment might have been performed. For example, the purpose might have been to prove the validity of the mathematical model. Alternatively, the purpose might have been to measure the parameters of the model. Yet another purpose might have been to develop the model so that it could be used to predict values of the dependent variable for combinations of the independent variables. To decide whether or not the experiment has been successful one usually considers the resulting accuracy. Is the accuracy of results sufficient to meet the criteria specified in the design phase? If not what can be done to improve the results?

1.2 Quantitative Experiments

The subject of this book is the design of quantitative experiments. We define quantitative experiments as experiments in which data is obtained for a dependent variable (or variables) as a function of an independent variable (or variables). The dependent and independent variables are then related through a mathematical model. These are experiments in which the variables are represented numerically.

One of the most famous quantitative experiments was performed by an Italian astronomer by the name of Giuseppe Piazzi of Palermo. In the late 1700's he set up an astronomical observatory which afforded him access to the southernmost sky in Europe at that time. His royal patron allowed him to travel to England where he supervised the construction of a telescope that permitted extremely accurate observations of heavenly bodies. Once the telescope was installed at the Royal Observatory in Palermo, Piazzi started work on a star catalog that was the most accurate that had been produced up to that time. In 1801 he stumbled across something in the sky that at first he thought was a comet. He took reading over a 42 day period when weather was permitting and then the object faded from view. Eventually it was recognized that the object was a planetoid in an orbit between Mars and Jupiter and he named it Ceres.

What made the discovery of Ceres such a memorable event in the history of science is the analysis that Gauss performed on the Piazzi data. To perform the analysis, Gauss developed the method of least squares which he described in his famous book ***Theoria Motus***. A translation of this book was published in English in 1857 [GA57]. By applying the method to calculate the orbit of Ceres, Gauss was able to accurately predict the reappearance of the planetoid in the sky. The method of least squares has been described as "the most nontrivial technique of modern statistics" [St86]. To this day, analysis of quantitative experiments relies to an overwhelming extent on this method. A simulation of Piazzi's experiment is included in Section 6.6 of this book.

1.3 Dealing with Uncertainty

The estimation of uncertainty is an integral part of data analysis. Typically uncertainty is expressed quantitatively as a value such as σ (the standard deviation) of whatever is being measured or computed. (The definition of σ is discussed below.) With every measurement we should include some indication regarding accuracy of the measurement. Some measurements are limited by the accuracy of the measuring instrument. For example, digital thermometers typically measure temperature to accuracies of 0.1°C. However, there are instruments that measure temperature to much greater accuracies. Alternatively, some measurements are error free but are subject to probability distributions. For example, consider a measurement of the number of people affected by a particular genetic problem in a group of 10000 people. If we examine all 10000 people and observe that twenty people test positive, what is our uncertainty? Obviously for this particular group of 10000 people there is no uncertainty in the recorded number. However, if we test a different group of 10000 people, the number that will test positive will probably be different than twenty. Can we make a statement regarding the accuracy of the number 20?

One method for obtaining an estimate of uncertainty is to repeat the measurement n times and record the measured values x_i, i = 1 to n. We can estimate σ (the standard deviation of the measurements) as follows:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}(x_i - x_{avg})^2 \qquad (1.3.1)$$

In this equation x_{avg} is the average value of the n measurements of x. A qualitative explanation for the need for n-1 in the denominator of this equation is best understood by considering the case in which only one measurement of x is made (i.e., $n = 1$). For this case we have no information regarding the "spread" in the measured values of x. A detailed derivation of this equation is included in most elementary books on statistics. The implication in Equation 1.3.1 is that we need to repeat measurements a number of times in order to obtain estimates of uncertainties. Fortunately this is rarely the case.

Often the instrument used to perform the measurement is provided with some estimation of the uncertainty of the measurements. Typically the estimation of σ is provided as a fixed percentage (e.g., σ = 1% of the value of x) or a fixed value (e.g., σ = 0.1°C). Sometimes the uncertainty is dependent upon the value of the quantity being measured in a more complex manner than just a fixed percentage or a constant value. For such cases the provider of the measuring instrument might supply this information in a graphical format or perhaps as an equation. For cases in which the data is calculated rather than measured, the calculation is incomplete unless it is accompanied by some estimate of uncertainty.

For measurements that are subject to statistical distributions like the example cited above regarding the genetic problem per 10000 people, often a knowledge of the distribution is sufficient to allow us to estimate the uncertainty. For that particular problem we could assume a Poisson distribution (discussed in Section 2.4) and the estimated value of σ is the square root of the number of people diagnosed as positive (i.e., $\sqrt{20} = 4.47$). We should note that our measurement is only accurate to a ratio of 4.47/20 which is approximately 22%. If we increase our sample size to 100,000 people and observe about 200 with the genetic problem, our measurement accuracy would be about $\sqrt{200} = 14.1$ which is approximately 7%. This is an improvement of more than a factor of 3 in fractional accuracy but at an increase in the cost for running the experiment by a factor of 10!

Once we have an estimation of σ, how do we interpret it? In addition to σ, we have a result (i.e., the value of whatever we are trying to determine) either from a measurement or from a calculation. Let us define the result as x and the true (but unknown value) of what we are trying to measure or compute as μ. Typically we assume that our best estimate of this true value of μ is x and that μ is located within a region around x. The size of the region is characterized by σ. In the preceding example our result x was 20 and the estimated value of σ was 4.47. This implies that if we were able to determine the true value of μ there is a "large" probability that it would be somewhere in the range 15 to 25. How "large" is a question that is considered in the discussion of probability distributions in Section 2.4. A typical assumption is that the probability of μ being greater or less than x is the same. In other words, our measurement or calculation includes a random error characterized by σ. Unfortunately this assumption is not always valid!

Sometimes our measurements or calculations are corrupted by **systematic errors**. Systematic errors are errors that cause us to either systematically under-estimate or over-estimate our measurements or computations. One source of systematic errors is an unsuccessful calibration of a measuring instrument. Another source is failure to take into consideration external factors that might affect the measurement or calculation (e.g., temperature effects). The choice of the mathematical model can also lead to systematic errors if it is overly simplistic or if it includes erroneous constants. Data analysis of quantitative experiments is based upon the assumption that the measured or calculated variables are not subject to systematic errors and that the mathematical model is a true representation of the process being modeled. If these assumptions are not valid, then errors are introduced into the results that do not show up in the computed values of the σ's. One can modify the least squares analysis to study the sensitivity of the results to systematic errors but whether or not systematic errors exist is a fundamental issue in any work of an experimental nature.

1.4 Parametric Models

Quantitative experiments are usually based upon parametric models. In this discussion we define **parametric models** as models utilizing a mathematical equation that describes the phenomenon under observation. As an example of a quantitative experiment based upon a parametric model, consider an experiment shown schematically in Figure 1.4.1. In this experiment a radioactive source is present and a detector is used to monitor radiation emanating from the source. Each time a radioactive particle enters the detector an "event" is noted. The recorder counts the number of events observed within a series of user specified time windows and thus produces a record of the number of counts observed as a function of time. The purpose of this experiment is to measure the half-life of a radioactive isotope. There is a single dependent variable ***counts*** (number of counts per unit of time) and a single independent variable ***time***. The parametric model for this particular experiment is:

$$counts = a_1 \cdot e^{-a_2 \cdot time} + a_3 \qquad (1.4.1)$$

This model has 3 parameters: the amplitude a_1, the decay constant a_2 and the background count rate a_3. The decay constant is related to the half-life (the time required for half of the isotope atoms to decay) as follows:

$$e^{-a_2 \cdot half\_life} = 1/2$$

$$half\_life = \frac{\ln(2)}{a_2} = \frac{0.69315}{a_2} \qquad (1.4.2)$$

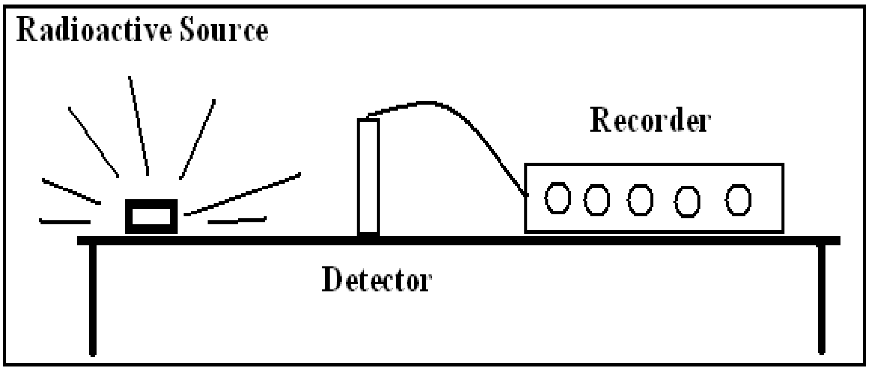

Figure 1.4.1 Experiment to Measure Half-life of a Radioisotope

The model equation (or equations) contains unknown parameters and the purpose of the experiment is often to determine the parameters including some indication regarding the accuracies (i.e., values of the σ's) of these parameters. There are many situations in which the values of the individual parameters are of no interest. All that is important for these cases is that the model can be used to predict values of the dependent variable (or variables) for other combinations of the independent variables. In addition, we are also interested in some measure of the accuracy (i.e.,σ) of the predictions.

We need to use mathematical terminology to define parametric models. Let us use the term y to denote the dependent variable and x to denote the independent variable. Usually y is a scalar, but when there is more than one dependent variable, y can denote a vector. The parametric model is the mathematical equation that defines the relationship between the dependent and independent variables. For the case of a single dependent and a single independent variable we can denote the model as:

$$y = f(x; a_1, a_2 .., a_p) \qquad (1.4.3)$$

The a_k's are the p unknown parameters of the model. The function f is based on either theoretical considerations or perhaps it is a function that seems to fit the measured values of y and x. Equation 1.4.1 is an example of a model in which p=3. The dependent variable y is ***counts*** and the independent variable x is ***time***.

When there is more than one independent variable, we can use the following equation to denote the model:

$$y = f(x_1, x_2 .., x_m; a_1, a_2 .., a_p) \qquad (1.4.4)$$

The x_j's are the m independent variables. As an example of an experiment in which there is more than one independent variable, consider an experiment based upon the layout shown in Figure 1.4.2. In this experiment a grid of detectors is embedded in a block of material. A source of radioactivity is placed near the block of material and the level of radioactivity is measured at each of the detectors. The count rate at each detector is a function of the position (x_1, x_2) within the block of material. The unknown parameters of the model are related to the radiation attenuation properties of the material.

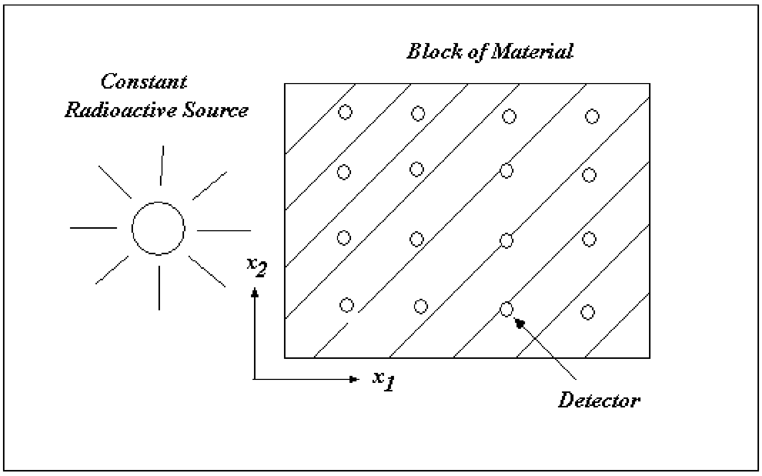

Figure 1.4.2 - Experiment to measure radioactive attenuation

If there is more than one dependent variable, we require a separate function for each element of the y vector:

$$y_l = f_l(x_1, x_2 .., x_m; a_1, a_2 .., a_p) \quad l = 1 \text{ to } d \qquad (1.4.5)$$

For cases of this type, y is a d dimensional vector and the subscript l refers to the l^{th} term of the y vector. It should be noted that some or all of the x_j's and the a_k's may be included in each of the d equations. The notation for the i^{th} data point for this l^{th} term of the y vector would be:

$$y_{l_i} = f_l(x_{1_i}, x_{2_i} .., x_{m_i}; a_1, a_2 .., a_p)$$

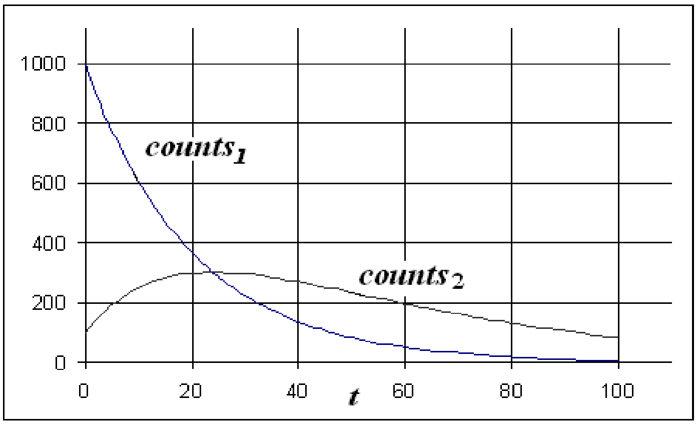

Figure 1.4.3 ***Counts* versus Time for Equations 1.4.6 and 1.4.7**
a_1=1000, a_2=100,* c_1*=0.05,* c_2*=0.025

An example of an experiment requiring a model of the form of Equation 1.4.5 can also be based upon the layout shown in Figure 1.4.1. If we assume that there are two radioactive species in the source and that species 2 is the daughter product of the species 1, we can measure the number of counts emanating from each species by discriminating the counts based upon the differing energies of the particles reaching the detector from the two species. Assuming that the two count rates ($\boldsymbol{counts_1}$ and $\boldsymbol{counts_2}$) are corrected for their background count rates, they are related to time $\boldsymbol{t}$ as follows:

$$counts_1 = a_1 \cdot e^{-c_1 \cdot t} \tag{1.4.6}$$

$$counts_2 = a_2 \cdot e^{-c_2 \cdot t} + a_1 \frac{c_2}{c_2 - c_1}\left(e^{-c_1 \cdot t} - e^{-c_2 \cdot t}\right) \tag{1.4.7}$$

For this example, there are four unknown parameters: the initial amplitudes $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$, and the decay constants $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$. A typical plot of $\boldsymbol{counts_1}$ and $\boldsymbol{counts_2}$ versus $\boldsymbol{t}$ is shown in Figure 1.4.3. A least squares analysis of the data will determine the values of these four parameters and estimates of their standard deviations.

A model is recursive if the functions defining the dependent variables $\boldsymbol{y_l}$ are interdependent. The form for the elements of recursive models is as follows:

$$y_l = f_l(x_1, x_2 .., x_m; y_1, y_2 .., y_d; a_1, a_2 .., a_p) \tag{1.4.8}$$

An example of a recursive model is the well-known prey-predator model of Kolmogorov [HO06, BR01, FR80]:

$$\frac{dy_1}{dt} = y_1 f_1(y_1, y_2) \tag{1.4.9}$$

$$\frac{dy_2}{dt} = y_2 f_2(y_1, y_2) \tag{1.4.10}$$

where $\boldsymbol{y_1}$ is the prey population and $\boldsymbol{y_2}$ is the predator population. The famous Italian mathematician Vito Volterra proposed a simple model to represent predator-prey interactions:

$$f_1 = a_1 - a_2 y_2 \tag{1.4.11}$$

$$f_2 = a_3 y_1 - a_4 \tag{1.4.12}$$

The parameter a_1 is the prey growth rate in the absence of predators and a_4 is the predator death rate in the absence of prey. The parameters a_2 and a_3 are the interaction coefficients. Increasing the predator population (i.e., y_2) causes a decrease in the prey population (i.e., y_1) and visa versa. Both of these equations are recursive: there is one independent variable t, four unknown parameters (a_1 to a_4) and two dependent variables (y_1 and y_2). We see that y_1 is dependent upon y_2 and y_2 is dependent upon y_1. The solution of Equations 1.4.9 and 1.4.10 introduces 2 new parameters: the initial values of y_1 and y_2 (i.e., y_{10} and y_{20}):

$$y_1 = y_{10} + \int y_1 f_1(y_1, y_2) dt \tag{1.4.13}$$

$$y_2 = y_{20} + \int y_2 f_2(y_1, y_2) dt \tag{1.4.14}$$

These two parameters can be treated as known constants or unknown parameters that are determined as part of the analysis of the data. Once a parametric model has been proposed and data is available, the task of data analysis must be performed. There are several possible objectives of interest to the analyst:

1) Compute the values of the p unknown parameters $a_1, a_2, \ldots a_p$.
2) Compute estimates of the standard deviations of the p unknown parameters.
3) Use the p unknown parameters to compute values of y for desired combinations of the independent variables $x_1, x_2, \ldots x_m$.
4) Compute estimates of the standard deviations σ_f for the values of $y = f(x)$ computed in 3.

It should be mentioned that the theoretically best solution to all of these objectives is achieved by applying the **method of maximum likelihood**. This method was proposed as a general method of estimation by the renowned statistician R. A. Fisher in the early part of the 20th century [e.g., FR92]. The method can be applied when the uncertainties associated with the observed or calculated data exhibit any type of distribution. However, when these uncertainties are normally distributed or when the normal distribution is a reasonable approximation, the method of maximum likelihood reduces to the **method of least squares** [WO06, HA01]. Fortunately, the assumption of normally distributed random errors is reasonable for most situations and thus the method of least squares is applicable for analysis of most quantitative experiments.

1.5 Basic Assumptions

The method of least squares can be applied to a wide variety of analyses of experimental data. The common denominator for this broad class of problems is the applicability of several basic assumptions. Before discussing these assumptions let us consider the measurement of a dependent variable Y_i. For the sake of simplicity, let us assume that the model describing the behavior of this dependent variable includes only a single independent variable. Using Equation 1.4.3 as the model that describes the relationship between x and y then y_i is the computed value of y at x_i. We define the difference between the measured and computed values as the residual R_i:

$$Y_i = y_i + R_i = f(x_i; a_1, a_2, \ldots a_p) + R_i \tag{1.5.1}$$

It should be understood that neither $\boldsymbol{Y_i}$ nor $\boldsymbol{y_i}$ are necessarily equal to the true value $\boldsymbol{\eta_i}$. In fact there might not be a single true value if the dependent variable can only be characterized by a distribution. However, for the sake of simplicity let us assume that for every value of $\boldsymbol{x_i}$ there is a unique true value (or a unique mean value) of the dependent variable that is $\boldsymbol{\eta_i}$. The difference between $\boldsymbol{Y_i}$ and $\boldsymbol{\eta_i}$ is the error (or uncertainty) $\boldsymbol{\varepsilon_i}$:

$$Y_i = \eta_i + \varepsilon_i \qquad (1.5.2)$$

The method of least squares is based upon the following assumptions:

1) If the measurement at $\boldsymbol{x_i}$ were to be repeated many times, then the values of error $\boldsymbol{\varepsilon_i}$ would be normally distributed with an average value of zero. Alternatively, if the errors are not normally distributed, the approximation of a normal distribution is reasonable.

2) The errors are uncorrelated. This is particularly important for time-dependent problems and implies that if a value measured at time $\boldsymbol{t_i}$ includes an error $\boldsymbol{\varepsilon_i}$ and at time $\boldsymbol{t_{i+k}}$ includes an error $\boldsymbol{\varepsilon_{i+k}}$ these errors are not related (i.e., uncorrelated). Similarly, if the independent variable is a measure of location, then the errors at nearby points are uncorrelated.

3) The standard deviations $\boldsymbol{\sigma_i}$ of the errors can vary from point to point. This assumption implies that $\boldsymbol{\sigma_i}$ is not necessarily equal to $\boldsymbol{\sigma_j}$.

The implication of the first assumption is that if the measurement of $\boldsymbol{Y_i}$ is repeated many times, the average value of $\boldsymbol{Y_i}$ would be the true (i.e., errorless) value $\boldsymbol{\eta_i}$. Furthermore, if the model is a true representation of the connection between $\boldsymbol{y}$ and $\boldsymbol{x}$ and if we knew the true values of the unknown parameters the residuals $\boldsymbol{R_i}$ would equal the errors $\boldsymbol{\varepsilon_i}$:

$$Y_i = \eta_i + \varepsilon_i = f(x_i; \alpha_1, \alpha_2, \ldots \alpha_p) + \varepsilon_i \qquad (1.5.3)$$

In this equation the true value of the $\boldsymbol{a_k}$ is represented as $\boldsymbol{\alpha_k}$. However, even if the measurements are perfect (i.e., $\boldsymbol{\varepsilon_i} = 0$), if $\boldsymbol{f}$ does not truly describe the dependency of $\boldsymbol{y}$ upon $\boldsymbol{x}$, then there will certainly be a difference between the measured and computed values of $\boldsymbol{y}$.

The first assumption of normally distributed errors is usually reasonable. Even if the data is characterized by other distributions (e.g., the binomial or Poisson distributions), the normal distribution is often a reasonable approximation. But there are problems where an assumption of normality causes improper conclusions. For example, in financial risk analysis the probability of catastrophic events (for example, the financial meltdown in the mortgage market in 2008) might be considerably greater than one might predict using normal distributions. To cite another area, earthquake predictions require analyses in which normal distributions cannot be assumed. Yet another area that is subject to similar problems is the modeling of insurance claims. Most of the data represents relatively small claims but there are usually a small fraction of claims that are much larger, negating the assumption of normality. Problems in which the assumption of normal error distributions is invalid are beyond the scope of this book. However, there is a large body of literature devoted to this subject. An extensive review of the subject is included in a book by Yakov Ben Haim [BE06].

One might ask when the second assumption (i.e., uncorrelated errors) is invalid. There are areas of science and engineering where this assumption is not really reasonable and therefore the method of least squares must be modified to take error correlation into consideration. Davidian and Giltinan discuss problem in the biostatistics field in which repeated data measurements are taken [DA95]. For example, in clinical trials, data might be taken for many different patients over a fixed time period. For such problems we can use the term Y_{ij} to represent the measurement at time t_i for patient j. Clearly it is reasonable to assume that ε_{ij} is correlated with the error at time t_{i+1} for the same patient. In this book, no attempt is made to treat such problems.

Many statistical textbooks include discussions of the method of least squares but use the assumption that all the σ_i's are equal. This assumption is really not necessary as the additional complexity of using varying σ_i's is minimal. Another simplifying assumption often used is that the models are linear with respect to the a_k's. This assumption allows a very simple mathematical solution but is too limiting for the analysis of many real-world experiments. This book treats the more general case in which the function f (or functions f_l) can be nonlinear.

1.6 Treatment of Systematic Errors

I did my graduate research in nuclear science at MIT. My thesis was a study of the fast fission effect in heavy water nuclear reactors and I was reviewing previous measurements [WO62]. The fast fission effect had been measured at two national laboratories and the numbers were curiously different. Based upon the values and quoted σ's, the numbers were many σ's apart. I discussed this with my thesis advisors and we agreed that one or both of the experiments was plagued by systematic errors that biased the results in a particular direction. We were proposing a new method which we felt was much less prone to systematic errors.

Results of experiments can often be misleading. When one sees a result stated as 17.3 ± 0.5 the reasonable assumption is that the true value should be somewhere within the range 16.8 to 17.8. Actually, if one can assume that the error is normally distributed about 17.3, and if 0.5 is the estimated standard deviation of the error, then the probability of the true value falling within the specified range is about 68%. The assumption of a normally distributed error centered at the measured value is based upon an assumption that the measurement is not effected by systematic errors. If, however, one can place an upper limit on all sources of systematic errors, then the estimated standard deviation can be modified to include treatment of systematic errors. By including systematic errors in the estimated standard deviation of the results, a more realistic estimate of the uncertainty of a measurement can be made.

As an example of an experiment based upon a single independent variable and a single dependent variable, consider the first experiment discussed in Section 1.4: the measurement of the half-life of a radioactive species. The mathematical model for the experiment is Equation 1.4.1 but let us assume that the background radiation is negligible (i.e., a_3 is close to zero). We could then use the following mathematical model:

$$counts = a_1 \cdot e^{-a_2 \cdot time} \qquad (1.6.1)$$

The fact that we have neglected to include a background term is a potential source of a systematic error. For this example, the source of the systematic error is the choice of the mathematical model itself. The magnitude of this systematic error can be estimated using the same computer program that is used to compute the values of the unknown parameters (i.e., a_1 and a_2).

Probably the greatest sources of systematic errors are measurement errors that introduce a bias in one direction or the other for the independent and dependent variables. One of the basic assumptions mentioned in the previous section is that the errors in the data are random about the true values. In other words, if a measurement is repeated n times, the average value would approach the true value as n becomes large. However, what happens if this assumption is not valid? When one can establish maximum possible values for such errors, the effect of these errors on the computed values of the parameters can be estimated by simple computer simulations. An example of this problem is included in Section 5.4.

Another source of systematic errors is due to errors in the values of parameters treated as known constants in the analysis of the data. As an example of this type of error consider the constants y_{10} and y_{20} in Equations 1.4.13 and 1.4.14. If these constants are treated as input quantities, then if there is uncertainty associated with their values, these uncertainties are a potential source of systematic errors. Estimates for the limits of these systematic errors can be made using the same software used for analysis of the data. All that needs to be done is to repeat the analysis for the range of possible values of the constants (e.g., y_{10} and y_{20}). This procedure provides a direct measurement of the effect of these sources of uncertainty on the resulting values of the unknown parameters (e.g., a_1 through a_4).

We can make some statements about combining estimates of systematic errors. Let us assume that we have identified ***nsys*** sources of systematic errors and that we can estimate the maximum size of each of these error sources. Let us define ε_{jk} as the systematic error in the measurement of a_j caused by the k^{th} source of systematic errors. The magnitude of the value of ε_j (the magnitude of the systematic error in the measurement of a_j caused by all sources) could range from zero to the sum of the absolute values of all the ε_{jk} 's. However, a more realistic estimate of ε_j is the following:

$$\varepsilon_j^2 = \sum_{k=1}^{k=nsys} \varepsilon_{jk}^2 \tag{1.6.2}$$

This equation is based upon the assumption that the systematic errors are uncorrelated. The total estimated uncertainty σ_j for the variable a_j should include the computed estimated standard deviation from the least squares analysis plus the estimated systematic error computed using Equation 1.6.2.

$$\sigma_j^2 = \sigma_{aj}^2 + \varepsilon_j^2 \tag{1.6.3}$$

Once, again this equation is based upon an assumption that the least squares estimated standard deviation and the estimated systematic error are uncorrelated.

1.7 Nonparametric Models

There are situations in which attempts to describe the phenomenon under observation by a single equation is extremely difficult if not impossible. For example, consider a dependent variable that is the future percentage return on stocks traded on the NYSE (New York Stock Exchange). One might be interested in trying to find a relationship between the future returns and several indicators that can be computed using currently available data. For this problem there is no underlying theory upon which a parametric model can be constructed. A typical approach to this problem is to use the historic data to define a surface and then use some sort of smoothing technique to make future predictions regarding the dependent variable [WO00].

The data plus the algorithm used to make the predictions are the major elements in what we define as a **nonparametric model.**

Nonparametric methods of data modeling predate the modern computer era [WO00]. In the 1920's two of the most well-known statisticians (Sir R. A. Fisher and E. S. Pearson) debated the value of such methods [HA90]. Fisher correctly pointed out that a parametric approach is inherently more efficient. Pearson was also correct in stating that if the true relationship between ***X*** and ***Y*** is unknown, then an erroneous specification in the function ***f*(*X*)** introduces a model bias that might be disastrous.

Hardle includes a number of examples of successful nonparametric models [HA90]. The most impressive is the relationship between change in height (cm/year) and age of women (Figure 1.7.1). A previously undetected growth spurt at around age 8 was noted when the data was modeled using a nonparametric smoother [GA84]. To measure such an effect using parametric techniques, one would have to anticipate this result and include a suitable term in ***f*(*X*)**.

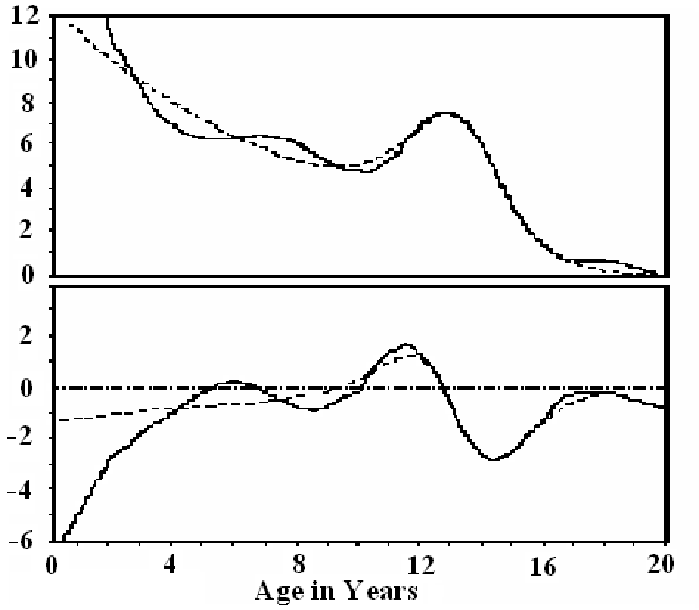

Figure 1.7.1 Human growth in women versus Age. The top graph is in cm/year. The bottom graph is acceleration in cm/year^2. The solid lines are from a model based upon nonparametric smoothing and the dashed lines are from a parametric fit [GA84, HA90].

The point at which one decides to give up attempts to develop a parametric model and cross over to nonparametric modeling is not obvious. For problems that are characterized by a large set of candidate predictors (i.e., predictors that might or might not be included in the final model) nonparametric modeling techniques can be used in an effort to seek out information rich subsets of the candidate predictor space. For example, when trying to model financial markets, one may consider hundreds of candidate predictors [WO00]. Financial market predictions are, of course, an area of intense world-wide interest. As a result there is considerable interest in applying nonparametric methods to the development of tools for making financial market predictions. A number of books devoted to this subject have been written in recent years (e.g., [AZ94, BA94, GA95, RE95, WO00, HO04, MC05]). If the nonparametric methods can successfully reduce the set of candidate predictors to a much smaller subset then a parametric approach to modeling might be possible. For such cases the design techniques considered in this book are applicable. However, if the experimental data will only be modeled using nonparametric techniques, the prediction analysis approach to design is not applicable.

1.8 Statistical Learning

The term **statistical learning** is used to cover a broad class of methods and problems that have become feasible as the power of the computer has grown. An in-depth survey of this field is covered in an excellent book by Hastie, Tibshirani and Friedman entitled ***The Elements of Statistical Learning: Data Mining, Inference and Prediction*** [HA01]. Their book covers both supervised and unsupervised learning. The goal of supervised learning is to predict an output variable as a function of a number of input variables (or as they are sometimes called: indicators or predictors). In unsupervised learning there is no particular output variable and one is interested in finding associations and patterns among the variables. The cornerstone of statistical learning is to *learn from the data*. The analyst has access to data and his or her goal is to make sense out of the available information.

Supervised learning problems can be subdivided into **regression** and **classification** problems. The goal in regression problems is to develop quantitative predictions for the dependent variable. The goal in classification problems is to develop methods for predicting to which class a particular data point belongs. An example of a regression problem is the development of a model for predicting the unemployment rate as a function of economic indictors. An example of a classification problem is the development of a model for predicting whether or not a particular email message is a spam message or a real message. In this book, the emphasis is on regression rather than classification problems. The design methods discussed in this book are not applicable for classification problems.