Data Analysis Using the Method of Least Squares

J. Wolberg

Data Analysis Using the Method of Least Squares

Extracting the Most Information from Experiments

With 58 Figures and 68 Tables

Springer

John Wolberg
Technion-Israel Institute of Technology
Faculty of Mechanical Engineering
32000 Haifa, Israel
E-mail: jwolber@attglobal.net

Library of Congress Control Number: 2005934230

ISBN-10 3-540-25674-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-25674-8 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media.

springer.com

Printed in Germany

Typesetting: Data prepared by the Author and by SPI Publisher Services
Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN 11010197 62/3141/SPI Publisher Services 5 4 3 2 1 0

For my parents, Sidney and Beatrice Wolberg ז"ל

My wife Laurie

My children and their families:

Beth, Gilad, Yoni and Maya Sassoon
David, Pazit and Sheli Wolberg
Danny, Iris, Noa, Adi and Liat Wolberg
Tamar, Ronen, Avigail and Aviv Kimchi

Preface

Measurements through quantitative experiments are one of the most fundamental tasks in all areas of science and technology. Astronomers analyze data from asteroid sightings to predict orbits. Computer scientists develop models for recognizing spam mail. Physicists measure properties of materials at low temperatures to understand superconductivity. Materials engineers study the reaction of materials to varying load levels to develop methods for prediction of failure. Chemical engineers consider reactions as functions of temperature and pressure. The list is endless. From the very small-scale work on DNA to the huge-scale study of black holes, quantitative experiments are performed and the data must be analyzed.

Probably the most popular method of analysis of the data associated with quantitative experiments is least squares. It has been said that the method of least squares was to statistics what calculus was to mathematics. Although the method is hardly mentioned in most engineering and science undergraduate curricula, many graduate students end up using the method to analyze the data gathered as part of their research. There is not a lot of available literature on the subject. Very few books deal with least squares at the level of detail that the subject deserves. Many books on statistics include a chapter on least squares but the treatment is usually limited to the simplest cases of linear least squares. The purpose of this book is to fill the gaps and include the type of information helpful to scientists and engineers interested in applying the method in their own special fields.

The purpose of many engineering and scientific experiments is to determine parameters based upon a mathematical model related to the phenomenon under observation. Even if the data is analyzed using least squares, the full power of the method is often overlooked. For example, the data can be weighted based upon the estimated errors associated with the data. Results from previous experiments or calculations can be combined with the least squares analysis to obtain improved estimate of the model parameters. In addition, the results can be used for predicting values of the dependent variable or variables and the associated uncertainties of the predictions as functions of the independent variables.

The introductory chapter (Chapter 1) includes a review of the basic statistical concepts that are used throughout the book. The method of least squares is developed in Chapter 2. The treatment includes development of mathematical models using both linear and nonlinear least squares. In Chapter 3 evaluation of models is considered. This chapter includes methods for measuring the "goodness of fit" of a model and methods for comparing different models. The subject of candidate predictors is discussed in Chapter 4. Often there are a number of candidate predictors and the task of the analyst is to try to extract a model using subspaces of the full candidate predictor space. In Chapter 5 attention is turned towards designing experiments that will eventually be analyzed using least squares. The subject considered in Chapter 6 is nonlinear least squares software. Kernel regression is introduced in the final chapter (Chapter 7). Kernel regression is a nonparametric modeling technique that utilizes local least squares estimates.

Although general purpose least squares software is available, the subject of least squares is simple enough so that many users of the method prefer to write their own routines. Often, the least squares analysis is a part of a larger program and it is useful to imbed it within the framework of the larger program. Throughout the book very simple examples are included so that the reader can test his or her own understanding of the subject. These examples are particularly useful for testing computer routines.

The REGRESS program has been used throughout the book as the primary least squares analysis tool. REGRESS is a general purpose nonlinear least squares program and I am its author. The program can be downloaded from *www.technion.ac.il/wolberg*.

I would like to thank David Aronson for the many discussions we have had over the years regarding the subject of data modeling. My first experiences with the development of general purpose nonlinear regression software were influenced by numerous conversations that I had with Marshall Rafal. Although a number of years have passed, I still am in contact with Marshall. Most of the examples included in the book were based upon software that I developed with Ronen Kimchi and Victor Leikehman and I would like to thank them for their advice and help. I would like to thank Ellad Tadmor for getting me involved in the research described in Section 7.7. Thanks to Richard Green for introducing me to the first English translation of Gauss's *Theoria Motus* in which Gauss developed the foundations of the method of least squares. I would also like to thank Donna Bossin for her help in editing the manuscript and teaching me some of the cryptic subtleties of WORD.

I have been teaching a graduate course on analysis and design of experiments and as a result have had many useful discussions with our students throughout the years. When I decided to write this book two years ago, I asked each student in the course to critically review a section in each chapter that had been written up to that point. Over 20 students in the spring of 2004 and over 20 students in the spring of 2005 submitted reviews that included many useful comments and ideas. A number of typos and errors were located as a result of their efforts and I really appreciated their help.

John R. Wolberg
Haifa, Israel
July, 2005

Contents

Chapter 1 INTRODUCTION

1.1 Quantitative Experiments

Most areas of science and engineering utilize **quantitative experiments** to determine parameters of interest. Quantitative experiments are characterized by measured variables, a mathematical model and unknown parameters. For most experiments the method of **least squares** is used to analyze the data in order to determine values for the unknown parameters.

As an example of a quantitative experiment, consider the following: measurement of the half-life of a radioactive isotope. Half-life is defined as the time required for the count rate of the isotope to decrease by one half. The experimental setup is shown in Figure 1.1.1. Measurements of ***Counts*** (i.e., the number of counts observed per time unit) are collected from time 0 to time ***tmax***. The mathematical model for this experiment is:

$$Counts = amplitude \cdot e^{-decay\_constant \cdot t} + background \qquad (1.1.1)$$

For this experiment, ***Counts*** is the **dependent variable** and time ***t*** is the **independent variable**. For this mathematical model there are 3 unknown parameters (***amplitude, decay_constant*** and ***background***). Possible sources of the background "noise" are cosmic radiation, noise in the instrumentation and sometimes a second much longer lived radioisotope within the source. The analysis will yield values for all three parameters but only the value of ***decay_constant*** is of interest. The half-life is determined from the resulting value of the decay constant:

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

$$half\_life = \frac{0.69315}{decay\_constant} \qquad (1.1.2)$$

The number 0.69315 is the natural logarithm of 2. This mathematical model is based upon the physical phenomenon being observed: the number of counts recorded per unit time from the radioactive isotope decreases exponentially to the point where all that is observable is the background noise.

There are alternative methods for conducting and analyzing this experiment. For example, the value of ***background*** could be measured in a separate experiment. One could then subtract this value from the observed values of ***Counts*** and then use a mathematical model with only two unknown parameters (***amplitude*** and ***decay_constant***):

$$Counts - background = amplitude \cdot e^{-decay\_constant \cdot t} \qquad (1.1.3)$$

The selection of a mathematical model for a particular experiment might be trivial or it might be the main thrust of the work. Indeed, the purpose of many experiments is to either prove or disprove a particular mathematical model. If, for example, a mathematical model is shown to agree with experimental results, it can then be used to make predictions of the dependent variable for other values of the independent variables.

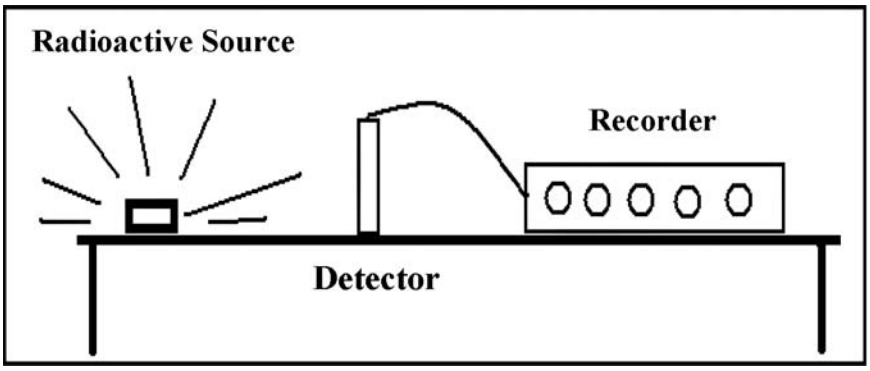

Figure 1.1.1 Experiment to Measure Half-life of a Radioisotope

Another important aspect of experimental work relates to the determination of the unknown parameters. Besides evaluation of these parameters by experiment, there might be an alternative calculation of the parameters based upon theoretical considerations. The purpose of the experiments for such cases is to confirm the theoretical results. Indeed, experiments go hand-in-hand with theory to improve our knowledge of the world around us.

Equations (1.1.1) and (1.1.3) are examples of mathematical models with only one independent variable (i.e., time ***t***) and only one dependent variable (i.e., ***Counts***). Often the mathematical model requires several independent variables and sometimes even several dependent variables. For example, consider classical chemical engineering experiments in which reaction rates are measured as functions of both pressure and temperature:

$$\boldsymbol{reaction\_rate = f(pressure, temperature)} \tag{1.1.4}$$

The actual form of the function ***f*** is dependent upon the type of reaction being studied.

The following example relates to an experiment that requires two dependent variables. This experiment is a variation of the experiment illustrated in Figure 1.1.1. Some radioactive isotopes decay into a second radioisotope. The decays from both isotopes give off signals of different energies and appropriate instrumentation can differentiate between the two different signals. We can thus measure count rates from each isotope simultaneously. If we call them ***c1*** and ***c2***, assuming background radiation is negligible, the appropriate mathematical model would be:

$$\boldsymbol{c1 = a1 \cdot e^{-d1 \cdot t}} \tag{1.1.5}$$

$$\boldsymbol{c2 = a2 \cdot e^{-d2 \cdot t} + a1 \frac{d2}{d2 - d1} \left(e^{-d1 \cdot t} - e^{-d2 \cdot t} \right)} \tag{1.1.6}$$

This model contains four unknown parameters: the two amplitudes (***a*1** and ***a*2**) and the two decay constants (***d*1** and ***d*2**). The two dependent variables are ***c*1** and ***c*2**, and the single independent variable is time ***t***. The time dependence of ***c*1** and ***c*2** are shown in Figure 1.1.2 for one set of the parameters.

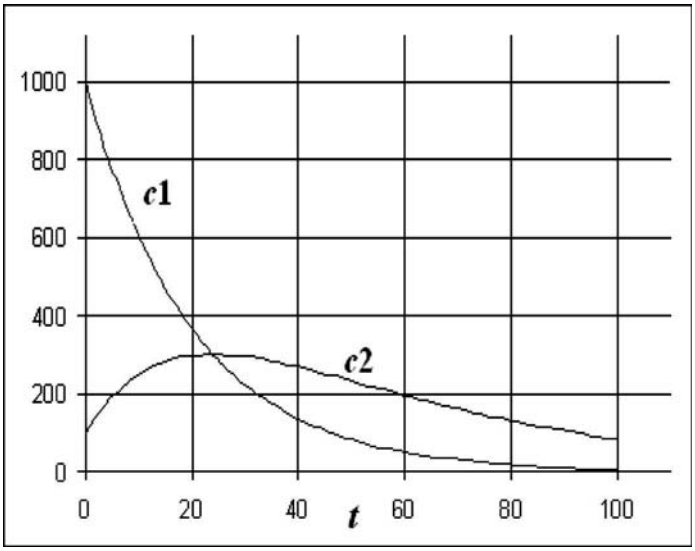

Figure 1.1.2 Counts versus Time for Equations 1.1.5 and 1.1.6
a*1=*1000*, *a*2=*100*, *d*1=*0.05*, *d*2=*0.025

The purpose of conducting experiments is not necessarily to prove or disprove a mathematical model or to determine parameters of a model. For some experiments the only purpose is to extract an equation from the data that can be used to predict values of the dependent variable (or variables) as a function of the independent variable (or variables). For such experiments the data is analyzed using different proposed equations (i.e., mathematical models) and the results are compared in order to select a "best" model.

We see that there are different reasons for performing quantitative experiments but what is common to all these experiments is the task of data analysis. In fact, there is no need to differentiate between physical experiments and experiments based upon computer generated data. Once data has been obtained, regardless of its origin, the task of data analysis commences. Whether or not the method of least squares is applicable depends upon the applicability of some basic assumptions. A discussion of the conditions allowing least squares analysis is included in Section 1.5: **Basic Assumptions.**

1.2 Dealing with Uncertainty

The estimation of uncertainty is an integral part of data analysis. It is not enough to just measure something. We always need an estimate of the accuracy of our measurements. For example, when we get on a scale in the morning, we know that the uncertainty is plus or minus a few hundred grams and this is considered acceptable. If, however, our scale were only accurate to plus or minus 10 kilograms this would be unacceptable. For other measurements of weight, an accuracy of a few hundred grams would be totally unacceptable. For example, if we wanted to purchase a gold bar, our accuracy requirements for the weight of the gold bar would be much more stringent. When performing quantitative experiments, we must take into consideration uncertainty in the input data. Also, the output of our analysis must include estimates of the uncertainty of the results. One of the most compelling reasons for using least squares analysis of data is that uncertainty estimates are obtained quite naturally as a part of the analysis. For almost all applications the standard deviation (σ) is the accepted measure of uncertainty. Let us say we need an estimate of the uncertainty associated with the measurement of the weight of gold bars. One method for obtaining such an estimate is to repeat the measurement n times and record the weights w_i , i = 1 to n. The estimate of σ (the estimated standard deviation of the weight measurement) is computed as follows:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{i=n} (w_i - w_{avg})^2 \qquad (1.2.1)$$

In this equation w_{avg} is the average value of the n measurements of w. The need for n-1 in the denominator of this equation is best explained by considering the case in which only one measurement of w is made (i.e., n = 1). For this case we have no information regarding the "spread" in the measured values of w.

Fortunately, for most measurements we don't have to estimate σ by repeating the measurement many times. Often the instrument used to perform the measurement is provided with some estimation of the accuracy of the measurements. Typically the estimation of σ is provided as a fixed percentage (e.g., σ= 1%) or a fixed value (e.g., σ= 0.5 grams). Sometimes the accuracy 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 measurement 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.

Once we have an estimation of σ, how do we interpret it? In addition to σ, we have a result either from measurements 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. 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). Data analysis of quantitative experiments is based upon the assumption that the measured or calculated independent and dependent variables are not subject to systematic errors. If this assumption is not true, 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.

1.3 Statistical Distributions

In nature most quantities that are observed are subject to a statistical distribution. The distribution is often inherent in the quantity being observed but might also be the result of errors introduced in the method of observation. An example of an inherent distribution can be seen in a study in which the percentage of smokers is to be determined. Let us say that one thousand people above the age of 18 are tested to see if they are smokers. The percentage is determined from the number of positive responses. It is obvious that if 1000 different people are tested the result will be different. If many groups of 1000 were tested we would be in a position to say some-

thing about the distribution of this percentage. But do we really need to test many groups? Knowledge of statistics can help us estimate the standard deviation of the distribution by just considering the first group!

As an example of a distribution caused by a measuring instrument, consider the measurement of temperature using a thermometer. Uncertainty can be introduced in several ways:

1) The persons observing the result of the thermometer can introduce uncertainty. If, for example, a nurse observes a temperature of a patient as 37.4°C, a second nurse might record the same measurement as 37.5°C. (Modern thermometers with digital outputs can eliminate this source of uncertainty.)

2) If two measurements are made but the time taken to allow the temperature to reach equilibrium is different, the results might be different. (Taking care that sufficient time is allotted for the measurement can eliminate this source of uncertainty.)

3) If two different thermometers are used, the instruments themselves might be the source of a difference in the results. This source of uncertainty is inherent in the quality of the thermometers. Clearly, the greater the accuracy, the higher is the quality of the instrument and usually, the greater the cost. It is far more expensive to measure a temperature to 0.001°C than 0.1°C!

We use the symbol $\boldsymbol{\Phi}$ to denote a distribution. Thus $\boldsymbol{\Phi(x)}$ is the distribution of some quantity $\boldsymbol{x}$. If $\boldsymbol{x}$ is a discrete variable then the definition of $\boldsymbol{\Phi(x)}$ is:

$$\sum_{xmin}^{xmax} \boldsymbol{\Phi(x)} = \mathbf{1} \tag{1.3.1}$$

If $\boldsymbol{x}$ is a continuous variable:

$$\int_{xmin}^{xmax} \boldsymbol{\Phi(x)dx} = \mathbf{1} \tag{1.3.2}$$

Two important characteristics of all distributions are the mean μ and the variance σ^2. The standard deviation σ is the square root of the variance. For discrete distributions they are defined as follows:

$$\mu = \sum_{xmin}^{xmax} x\Phi(x) \tag{1.3.3}$$

$$\sigma^2 = \sum_{xmin}^{xmax} (x-\mu)^2\Phi(x) \tag{1.3.4}$$

For continuous distributions:

$$\mu = \int_{xmin}^{xmax} x\Phi(x)dx \tag{1.3.5}$$

$$\sigma^2 = \int_{xmin}^{xmax} (x-\mu)^2\Phi(x)\,dx \tag{1.3.6}$$

The normal distribution

When x is a continuous variable the normal distribution is often applicable. The normal distribution assumes that the range of x is from -∞ to ∞ and that the distribution is symmetric about the mean value μ. These assumptions are often reasonable even for distributions of discrete variables, and thus the normal distribution can be used for some distributions of discrete variables. The equation for a normal distribution is:

$$\Phi(x) = \frac{1}{\sigma(2\pi)^{1/2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1.3.7}$$

The normal distribution is shown in Figure 1.3.1 for various values of the standard deviation σ. We often use the term **standard normal distribution** to characterize one particular distribution: a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The symbol u is usually used to denote this distribution. Any normal distribution can be transformed into a standard normal distribution by subtracting μ from the values of x and then dividing this difference by σ.

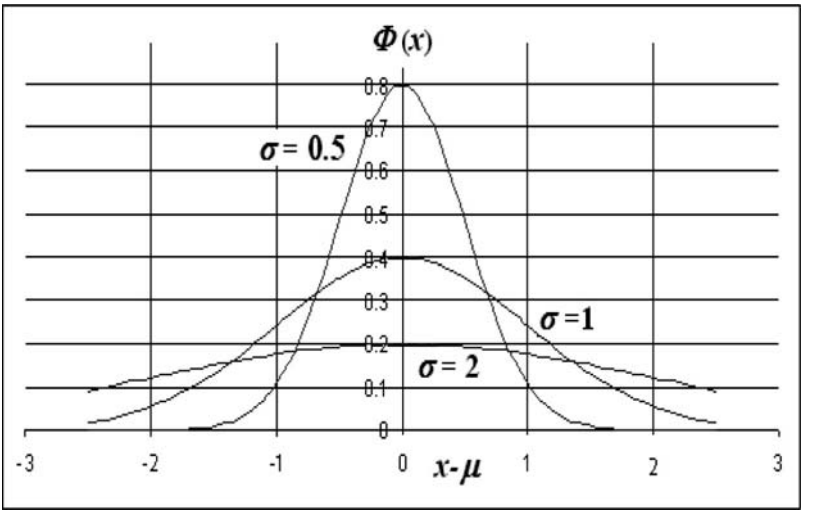

Figure 1.3.1 $\Phi(x)$ **vs** $x-\mu$ **for Normal Distribution (**σ**=0.5, 1 and 2).**

We can define the **effective range** of the distribution as the range in which a specified percentage of the data can be expected to fall. If we specify the effective range of the distribution as the range between $\mu \pm \sigma$, then 68.3% of all measurements would fall within this range. Extending the range to $\mu \pm 2\sigma$, 95.4% would fall within this range and 99.7% would fall within the range $\mu \pm 3\sigma$. The true range of any normal distribution is always $-\infty$ to ∞. Values of the percentage that fall within 0 to u (i.e., $(x-\mu)/\sigma$) are included in tables in many sources [e.g., AB64, FR92]. The standard normal table is also available online [ST03]. Approximate equations corresponding to a given value of probability are also available (e.g., See Appendix B).

The normal distribution is not applicable for all distributions of continuous variables. In particular, if the variable x can only assume positive values and if the mean of the distribution μ is close to zero, then the normal distribution might lead to erroneous conclusions. If however, the value of μ is large (i.e., $\mu/\sigma >> 1$) then the normal distribution is usually a good approximation even if negative values of x are impossible.

We are often interested in understanding how the mean of a sample of n values of x (i.e., x_{avg}) is distributed. It can be shown that the standard deviation of the value of x_{avg} has a standard deviation of $\sigma/\sqrt{n}$. Thus the quantity $(x_{avg}-\mu) / (\sigma/\sqrt{n})$ follows the standard normal distribution u. For

example, let us consider a population with a mean value of 50 and a standard deviation of 10. If we take a sample of $n = 100$ observations and then compute the mean of this sample, we would expect that this mean would fall in the range 49 to 51 with a probability of about 68%. In other words, even though the population σ is 10, the standard deviation of an average of 100 observations is only $10/\sqrt{100} = 1$.

The binomial distribution

When x is a discrete variable of values 0 to n (where n is a relatively small number), the binomial distribution is usually applicable. The variable x is used to characterize the number of **successes** in n trials where p is the probability of a single success for a single trial. The symbol $\Phi(x)$ is thus the probability of obtaining exactly x successes. The number of successes can theoretically range from 0 to n. The equation for the distribution is:

$$\Phi(x) = \frac{n!}{x!\,(n-x)!}\, p^x (1-p)^{n-x} \tag{1.3.8}$$

As an example, consider the following problem: what is the probability of drawing the Ace of Spades from a deck of cards if the total number of trials is 3. After each trial the card drawn is reinserted into the deck and the deck is shuffled. For this problem the possible values of x are 0, 1, 2 and 3. The value of p is 1/52 as there are 52 different cards in a deck: the Ace of Spades and 51 other cards. The probability of not drawing the Ace of Spades in any of the 3 trials is:

$$\Phi(0) = \frac{3!}{0!\,(3)!}\, p^0 (1-p)^3 = (51/52)^3 = 0.9434$$

The probability of drawing the Ace of Spades once is:

$$\Phi(1) = \frac{3!}{1!\,(2)!}\, p^1 (1-p)^2 = \frac{6}{2}(1/52)^1 (51/52)^2 = 0.0555$$

The probability of drawing the Ace of Spades twice is:

$$\Phi(2) = \frac{3!}{2!\,(1)!}\,p^2(1-p)^1 = \frac{6}{2}(1/52)^2(51/52)^1 = 0.00109$$

The probability of drawing the Ace of Spades all three times is:

$$\Phi(3) = \frac{3!}{3!\,(0)!}\,p^3(1-p)^0 = (1/52)^3 = 0.000007$$

The sum of all 4 of these probable outcomes is one. The probability of drawing the Ace of Spades at least once is 1 - 0.9434 = 0.0566.

The mean value μ and standard deviation σ of the binomial distribution can be computed from the values of n and p:

$$\mu = np \tag{1.3.9}$$

$$\sigma = (np(1-p))^{1/2} \tag{1.3.10}$$

Equation 1.3.9 is quite obvious. If, for example, we flip a coin 100 times, what is the average value of the number of heads we would observe? For this problem, p = ½, so we would expect to see on average 100 * 1/2 = 50 heads. The equation for the standard deviation is not obvious, however the proof of this equation can be found in many elementary textbooks on statistics. For this example we compute σ as $(100*1/2*1/2)^{1/2}$ = 5. Using the fact that the binomial distribution approaches a normal distribution for values of $\mu >> 1$, we can estimate that if the experiment is repeated many times, the numbers of heads observed will fall within the range 45 to 55 about 68% of the time.

The Poisson distribution

The binomial distribution (i.e., Equation 1.3.8) becomes unwieldy for large values of n. The Poisson distribution is used for a discrete variable x that can vary from 0 to ∞. If we assume that we know the mean value μ of the distribution, then $\Phi(x)$ is computed as:

$$\Phi(x) = \frac{e^{-\mu}\mu^x}{x!} \tag{1.3.11}$$

It can be shown that the standard deviation σ of the Poisson distribution is:

$$\sigma = \mu^{1/2} \tag{1.3.12}$$

If μ is a large value, the normal distribution is an excellent approximation of a Poisson distribution.

As an example of a Poisson distribution, consider the observation of a rare genetic problem. Let us assume that the problem is observed on average 2.3 times per 10000 people. For practical purposes n is close to ∞ so we can assume that the Poisson distribution is applicable. We can compute the probability of observing x people with the genetic problem out of a sample population of 10000 people. The probability of observing no one with the problem is:

$$\Phi(0) = e^{-2.3}\, 2.3^0 / 0! = e^{-2.3} = 0.1003$$

The probability of observing one person with the problem is:

$$\Phi(1) = e^{-2.3}\, 2.3^1 / 1! = 2.3e^{-2.3} = 0.2306$$

The probability of observing two people with the problem is:

$$\Phi(2) = e^{-2.3}\, 2.3^2 / 2! = 2.3^2 e^{-2.3} / 2 = 0.2652$$

The probability of observing three people with the problem is:

$$\Phi(3) = e^{-2.3}\, 2.3^3 / 3! = 2.3^3 e^{-2.3} / 6 = 0.2136$$

From this point on, the probability $\Phi(x)$ decreases more and more rapidly and for all intents and purposes approaches zero for large values of x.

Another application of Poisson statistics is for counting experiments in which the number of counts is large. For example, consider observation of a radioisotope by an instrument that counts the number of signals emanating from the radioactive source per unit of time. Let us say that 10000 counts are observed. Our first assumption is that 10000 is our best esti-

mate of the mean μ of the distribution. From equation 1.3.12 we can then estimate the standard deviation σ of the distribution as $10000^{1/2} = 100$. In other words, in a counting experiment in which 10000 counts are observed, the accuracy of this observed count rate is approximately 1% (i.e., 100/10000 = 0.01). To achieve an accuracy of 0.5% we can compute the required number of counts:

$$0.005 = \sigma / \mu = \mu^{1.2} / \mu = \mu^{-1/2}$$

Solving this equation we get a value of $\mu = 40000$. In other words to double our accuracy (i.e., halve the value of σ) we must increase the observed number of counts by a factor of 4.

The χ^2 distribution

The χ^2 (chi-squared) distribution is defined using a variable u that is normally distributed with a mean of 0 and a standard deviation of 1. This u distribution is called the standard normal distribution. The variable $\chi^2(k)$ is called the χ^2 value with k degrees of freedom and is defined as follows:

$$\chi^2(k) = \sum_{i=1}^{i=k} u_i^2 \tag{1.3.13}$$

In other words, if k samples are extracted from a standard normal distribution, the value of $\chi^2(k)$ is the sum of the squares of the u values. The distribution of these values of $\chi^2(k)$ is a complicated function:

$$\Phi(\chi^2(k)) = \frac{(\chi^2)^{k/2-1}}{2^{k/2}\Gamma(k/2)} \exp(-\chi^2/2) \tag{1.3.14}$$

In this equation Γ is called the gamma function and is defined as follows:

$$\Gamma(k/2) = (k/2-1)(k/2-2)...3*2*1 \quad \textit{for k even}$$
$$\Gamma(k/2) = (k/2-1)(k/2-2)...3/2*1/2*\pi^{1/2} \quad \textit{for k odd} \tag{1.3.15}$$

Equation 1.3.14 is complicated and rarely used. Of much greater interest is determination of a range of values from this distribution. What we are

more interested in knowing is the probability of observing a value of χ^2 from 0 to some specified value. This probability can be computed from the following equation [AB64]:

$$P(\chi^2/k) = \frac{1}{2^{k/2}\,\Gamma(k/2)} \int_0^{\chi^2} t^{k/2-1} e^{-t/2}\, dt \qquad (1.3.16)$$

For small values of k (typically up to k=30) values of χ^2 are presented in a tabular format [e.g., AB64, FR92, ST03] but for larger values of k, approximate values can be computed (using the normal distribution approximation described below). The tables are usually presented in an inverse format (i.e., for a given value of k, the values of χ^2 corresponding to various probability levels are tabulated). As an example of the use of this distribution, let us consider an experiment in which we are testing a process to check if something has changed. Some variable x characterizes the process. We know from experience that the mean of the distribution of x is μ and the standard deviation is σ. The experiment consists of measuring 10 values of x. An initial check of the computed average value for the 10 values of x is seen to be close to the historical value of μ but can we make a statement regarding the variance in the data? We would expect that the following variable would be distributed as a standard normal distribution (μ=0, σ=1):

$$u = \frac{(x-\mu)}{\sigma} \qquad (1.3.17)$$

Using Equation 1.3.17, 1.3.13 and the 10 values of x we can compute a value for χ^2. Let us say that the value obtained is 27.2. The question that we would like to answer is what is the probability of obtaining this value or a greater value by chance? From [ST03] it can be seen that for k = 10, there is a probability of 0.5% that the value of χ^2 will exceed 25.188. (Note that the value of k used was 10 and not 9 because the historical value of μ was used in Equation 1.3.17 and not the mean value of the 10 observations.) The value observed (i.e., 27.2) is thus on the high end of what we might expect by chance and therefore some problem might have arisen regarding the process under observation.

Two very useful properties of the χ^2 distribution are the mean and standard deviation of the distribution. For k degrees of freedom, the mean is k and the standard deviation is $\sqrt{2k}$. For large values of k, we can use the fact

that this distribution approaches a normal distribution and thus we can easily compute ranges. For example, if k = 100, what is the value of χ^2 for which only 1% of all samples would exceed it by chance? For a standard normal distribution, the 1% limit is 2.326. The value for the χ^2 distribution would thus be $\mu + 2.326*\sigma = k + 2.326*(2k)^{1/2} = 100 + 31.2 = 131.2$.

An important use for the χ^2 distribution is analysis of variance. The **variance** is defined as the standard deviation squared. We can get an **unbiased estimate** of the variance of a variable x by using n observations of the variable. Calling this unbiased estimate as s^2, we compute it as follows:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x - x_{avg})^2 \tag{1.3.18}$$

The quantity $(n\text{-}1)s^2/\sigma^2$ is distributed as χ^2 with n-1 degrees of freedom. This fact is fundamental for least squares analysis.

The *t* distribution

The t distribution (sometimes called the student-t distribution) is used for samples in which the standard deviation is not known. Using n observations of a variable x, the mean value x_{avg} and the unbiased estimate s of the standard deviation can be computed. The variable t is defined as:

$$t = (x_{avg} - \mu)/(s/\sqrt{n}) \tag{1.3.19}$$

The t distribution was derived to explain how this quantity is distributed. In our discussion of the normal distribution, it was noted that the quantity $(x_{avg}\text{-}\mu) / (\sigma/\sqrt{n})$ follows the standard normal distribution u. When σ of the distribution is not known, the best that we can do is use s instead. For large values of n the value of s approaches the true value of σ of the distribution and thus t approaches a standard normal distribution. The mathematical form for the t distribution is based upon the observation that Equation 1.3.19 can be rewritten as:

$$t = \frac{(x_{avg} - \mu)}{(\sigma / \sqrt{n})}(\sigma / s) \tag{1.3.20}$$

The term σ/s is distributed as $((n\text{-}1) / \chi^2)^{1/2}$ where χ^2 has n-1 degrees of freedom. Thus the mathematical form of the t distribution is derived from the product of the standard normal distribution and $((n\text{-}1)/ \chi^2(n\text{-}1))^{1/2}$. Values of t for various percentage levels for n-1 up to 30 are included in tables in many sources [e.g., AB64, FR92]. The t table is also available online [ST03]. For values of $n > 30$, the t distribution is very close to the standard normal distribution.

For small values of n the use of the t distribution instead of the standard normal distribution is necessary to get realistic estimates of ranges. For example, consider the case of 4 observations of x in which x_{avg} and s of the measurements are 50 and 10. The value of $s / \sqrt{n}$ is 5. The value of t for n - 1 = 3 degrees of freedom and 1% is 4.541. We can use these numbers to determine a range for the true (but unknown value) of μ:

$$27.30 = 50 - 4.541 * 5 <= \mu <= 50 + 4.541 * 5 = 77.71$$

In other words, the probability of μ being below 27.30 is 1%, above 77.71 is 1% and within this range is 98%. Note that the value of 4.541 is considerably larger than the equivalent value of 2.326 for the standard normal distribution. It should be noted, however, that the t distribution approaches the standard normal rather rapidly. For example, the 1% limit is 2.764 for 10 degrees of freedom and 2.485 for 25 degrees of freedom. These values are only 19% and 7% above the standard normal 1% limit of 2.326.

The *F* distribution

The F distribution plays an important role in data analysis. This distribution was named to honor R.A. Fisher, one of the great statisticians of the 20th century. The F distribution is defined as the ratio of two χ^2 distributions divided by their degrees of freedom:

$$F = \frac{\chi^2(k_1) / k_1}{\chi^2(k_2) / k_2} \tag{1.3.21}$$