Statistics for Analysis of Experimental Data

Catherine A. Peters

Department of Civil and Environmental Engineering
Princeton University
Princeton, NJ 08544

Published as a chapter in the

Environmental Engineering Processes Laboratory Manual

S. E. Powers, Ed.

AEESP, Champaign, IL

2001

Statistics for Analysis of Experimental Data

Catherine A. Peters

Department of Civil and Environmental Engineering
Princeton University
Princeton, NJ 08544

Statistics is a mathematical tool for quantitative analysis of data, and as such it serves as the means by which we extract useful information from data. In this chapter we are concerned with data that are generated via experimental measurement. Experimentation often generates multiple measurements of the same thing, i.e. replicate measurements, and these measurements are subject to error. Statistical analysis can be used to summarize those observations by estimating the average, which provides an estimate of the true mean. Another important statistical calculation for summarizing the observations is the estimate of the variance, which quantifies the uncertainty in the measured variable. Sometimes we have made measurements of one quantity and we want to use those measurements to infer values of a derived quantity. Statistical analysis can be used to propagate the measurement error through a mathematical model to estimate the error in the derived quantity. Sometimes we have measured two different things and we want to know whether there really is a difference between the two measured values. Analysis of variance (*t*-tests) can be used to estimate the probability that the underlying phenomena are truly different. Finally, we may have measured one variable under a variety of conditions with regard to a second variable. Regression analysis can be used to come up with a mathematical expression for the relationship between the two variables. These are but a few of the many applications of statistics for analysis of experimental data. This chapter presents a brief overview of these applications in the context of typical experimental measurements in the field of environmental engineering.

This chapter is necessarily brief in presentation. Students who seek a deeper understanding of these principles should study a textbook on statistical analysis of experimental data. The bibliography at the end of this chapter lists some useful textbooks, some of which are directly aimed at environmental engineers and scientists.

Error Analysis and Error Propagation

Errors in Measured Quantities and Sample Statistics

A very important thing to keep in mind when learning how to design experiments and collect experimental data is that our ability to observe the real world is not perfect. The observations we make are never exactly representative of the process we think we are observing. Mathematically, this is conceptualized as:

$$\text{measured value} = \text{true value} \pm \text{error} \tag{1}$$

The error is a combined measure of the inherent variation in the phenomenon we are observing and the numerous factors that interfere with the measurement. Every effort should be made to reduce systematic errors through efforts such as calibration of measurement instruments. It is impossible to totally eliminate all measurement error. If the underlying error is truly random (not biased) then we can still gain useful information by making multiple observations (i.e. **replicates**) and calculating the average. In order for the sample to be truly representative of the underlying phenomenon that is being measured it must be a

random sample. For example, let's say that you are running an experiment in which you have set up eight batch reactors and you plan to sacrifice one batch reactor every hour to measure the concentration of some chemical. Every time you select a batch reactor you should randomly select from the remaining reactors. You should not sample the reactors in the same order as you prepared them nor should you sample the reactors in the order in which they are positioned on your bench top. You never know how these other factors may influence the controlling processes in the reactors. By randomly sampling the reactors, any systematic error due to other factors is randomly distributed across your measurements. Randomness helps to ensure **independence** of the observations. When we say that we want "independent observations" what we really mean is that we want the errors in the observations to be independent of each other. Aside from nonrandom sampling, there are other laboratory activities that could jeopardize independence of the observations. For example, if an inexperienced experimentalist gets better at making a certain type of measurement, then the error may get smaller over time. In this case, the error is a function of the order in which the measurement is made and the errors are not independent. Similarly, if a measurement device wears out every time it is used then the error may increase over time. This too would produce errors that are not independent. Random sampling and other efforts to make the observation errors independent help to ensure representativeness. If all the observations are truly **representative** of the same underlying phenomenon, then they all have the same mean and variance, i.e. the errors are **identically distributed**. Sometimes the acronym IID is used to collectively refer to the criteria that a sample of observations is independent (I) and identically distributed (ID).

Given a sample of n observations, the **sample average** is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

where x_i represents the ith individual observation. The sample average is a **statistic** that is an estimate of η, the **mean**, or central tendency, of the underlying **random variable**. The **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{3}$$

The sample variance is a **statistic** that is an estimate of the **variance**, σ^2, in the underlying **random variable**. Another useful statistic is the **sample standard deviation**, s, which is the square root of the sample variance, σ. The quantity n-1 is the number of degrees of freedom associated with the sample standard deviation.

It is often the case that we are more interested in the estimate of the mean than in the individual observations. What we really want to know then is what is the variance in the average value. That is, how does the variance in x translate into uncertainty in our ability to estimate the mean? The **standard error of the mean** is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{4}$$

which also has n-1 degrees of freedom. Clearly, when the number of observations, n, is large, the uncertainty in the estimate of the mean is small. This relationship demonstrates that there is more uncertainty in an individual observation than in the estimated mean. Even if the underlying phenomenon

is quite variable and there are significant measurement errors, it is still possible to reduce uncertainty in the estimate of the mean by making many measurements.

> EXAMPLE. A student collects a series of twelve groundwater samples from a well. To start, she measures the dissolved oxygen concentration in six of these. Her observations in mg/L are:
>
> 8.8, 3.1, 4.2, 6.2, 7.6, 3.6
>
> The sample average is 5.6 mg/L. The sample standard deviation is 2.3 mg/L. This value can be interpreted as the error, or uncertainty, in any given measurement of the dissolved oxygen concentration. Note that the variation in these data represent both natural variation in the oxygen concentration in the water as well as variation due to measurement error. The standard error of the mean is 0.95 mg/L. Notice that this is considerably smaller than the sample standard deviation. After examining these statistics the student decides that the uncertainty in the estimate of the mean is unacceptably large. She proceeds to measure the dissolved oxygen concentration in each of the six remaining samples. The additional observations in mg/L are:
>
> 5.2, 8.6, 6.3, 1.8, 6.8, 3.9
>
> The grand average of all twelve observations is 5.5 mg/L and the standard deviation of the sample of twelve observations is 2.2 mg/L. These statistics are comparable to those of the smaller data set which provides some evidence that the original six observations are representative of the underlying phenomenon. The new standard error of the mean is 0.65 mg/L. The reduction in the uncertainty in the estimate of the mean results from having a larger number of observations in the sample.

The Normal Distribution

It is very often the case that an experimentalist will use a calculated sample average and standard error to infer something about the probability of the random variable under observation or its relationship to other random variables. To do this one must make an assumption about the shape of the probability distribution of the errors in the experimental measurements. Most statistical techniques require an assumption that the measurement errors have a **normal probability distribution**. The normal distribution is also frequently called the **Gaussian** distribution. A plot of a **probability distribution function** (PDF) for a normally distributed random variable x with mean of zero and standard deviation of unity is shown in Figure 1a. For a given value of x, the value on the y axis is $f(x)$, the probability density. The normal PDF is symmetric, centered at the mean of x, and it extends from negative infinity to positive infinity. By definition, the area under any probability distribution function equals unity. For a normal probability distribution, 68% of the area under the curve lies within $\eta \pm \sigma$, meaning that 68% of the total probability is within one standard deviation of the mean. Practically speaking, one would expect that roughly 2/3 of one's observations would fall within this range. The area under the curve within $\eta \pm 2\sigma$ captures 95% of the total probability and the area under the curve within $\eta \pm 3\sigma$ captures 99.7% of the total probability. Another way to view the normal distribution is as a **cumulative distribution function** (CDF), shown in Figure 1b. For a given value of x, the value on the y axis, $F(x)$, is the cumulative probability associated with values of the random variable less than or equal to x.

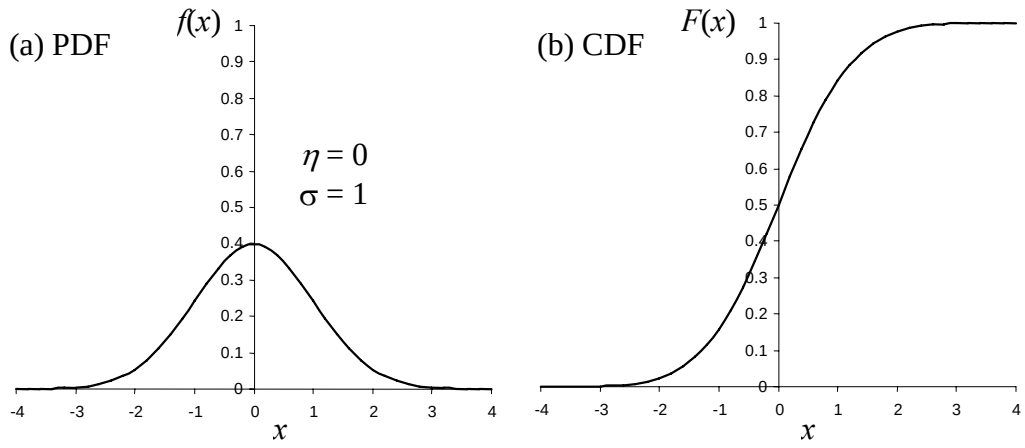


Figure 1a and 1b. The PDF of the normal probability distribution (a) and the CDF of the normal probability distribution (b) for a random variable x with mean of zero and standard deviation of unity.

Note that the stipulation for application of many statistical techniques is that the errors in the observations are normally distributed and not that the random variable itself is normally distributed. This is an important distinction because many environmental variables have distributions other than the normal distribution. For example, random variables that cannot assume negative values such as concentrations and random variables that vary over orders of magnitude such as hydraulic conductivity of a porous medium are typically lognormally distributed, i.e. the logarithm of the random variable is normally distributed. Another positively skewed probability distribution function that is widely used to describe environmental variables is the gamma distribution. For example, precipitation rates are often described using the gamma distribution.

If numerous replications have been made for a given measurement, then it is possible to examine whether the observations have a normally distributed error structure. This is typically done using a normal probability plot. A **normal probability plot** is a form of the normal CDF in which the y axis has been modified such that the cumulative distribution function appears to be linear. One can generate such a plot by constructing a rank-ordered list of the observations, estimating the cumulate probabilities, and plotting on special graph paper called normal probability paper (for more details see McBean and Rovers, 1998).

Despite the fact that a random variable may vary over time and space according to a non-normal probability distribution, it is quite possible that the observation of a particular value of a non-normally distributed random variable can be described using the normal distribution. As stated earlier, the error represents both variation inherent in the random variable as well as measurement error. The latter arises due to numerous small factors related to experimental design, sampling, detection, and analysis. For example, the manufacturer of a thermometer may not have calibrated it very well so the temperature markings do not exactly match the corresponding level of mercury. There may be impurities in the mercury so that it does not expand and contract in a reproducible fashion. The experimentalist's line of sight causes parallax error in reading the markings. The temperature varies slightly over the time period of the measurement. The temperature varies spatially in the region where the experimentalist would like to record the temperature. Collectively, these errors add up to generate the imprecision in an experimental measurement. The **central limit theorem** says that as the number of variables in a sum increases the distribution of the sum of random variables approaches the normal distribution regardless of the shape of the distribution of the individual random variables. Experimental measurement error is the aggregate of a large number of contributing errors. If the sources of error are numerous (as they usually are) then by the central limit theorem we can say that experimental errors tend to have a normal distribution. Furthermore,

often we use statistical tools to make an inference about a sample average, which is a further summation of values that are themselves likely to have normally distributed errors. This provides additional justification to assume that sample averages have normally distributed errors.

It is often the case in experimental studies that we don't have enough observations to generate a normal probability plot and make judgments about the shape of the underlying probability distribution. In the example above, even with as many as twelve observations it may be difficult to judge the linearity of a normal probability plot. Often we make only two or three replicate measurements which makes it impossible to construct a meaningful probability plot. In these cases, we have to assume a normally distributed error structure. Fortunately, the central limit theorem provides a theoretical basis for making this assumption in experimental measurements.

Confidence Intervals

For any estimated statistic, such as a sample average, for which we have an estimated value and an estimate of the standard error in that statistic, we can report **confidence intervals**. If the errors in the measurement variable, x, have a normal probability distribution and if the observations are independent, then the probability distribution for the error in the sample average, normalized by the standard error in the sample average, is the ***t*-distribution**. The t-distribution is a symmetric probability distribution centered at zero, like the normal probability distribution. The difference is that the t-distribution has a variance that depends on the degrees of freedom of the standard error in the statistic of interest. Recall that $s_{\bar{x}}$ has n-1 degrees of freedom. If very few measurements have been taken, the number of degrees of freedom is very small and the t-distribution has a very large variance.

The t-distribution is used to determine a t-statistic which is then used to calculate a confidence interval for the true value of the mean, η. The t-statistic of interest is that which bounds a chosen level of probability, 1-α, for the t-distribution with n-1 degrees of freedom. For example at a 90% probability level 1-α = 0.90 and α = 0.10. Most statistics textbooks have tables of values of the t-statistic for various levels of probability and values of degrees of freedom, and statistical software packages can compute t-statistics. The 1-α confidence interval for η is

$$\bar{x} \pm t_{n-1;\alpha/2} s_{\bar{x}} \qquad (5)$$

The reason the appropriate t-statistic is that which corresponds to the $\alpha/2$ probability level is because that value represents one side of a symmetric two-sided interval. We say that there is a 1-α probability that the confidence interval contains the true value of η. Conventionally used probability levels are the 90% (somewhat confident), 95% (fairly confident), and 99% (quite confident) probability levels.

EXAMPLE. For the oxygen concentration data discussed in the example above, what is the 95% confidence interval for the mean? The standard error of the mean has 12-1=11 degrees of freedom. The t-statistic that corresponds to the t-distribution for 11 degrees of freedom and 95% probability level is

$$t_{11;\,0.025} = 2.201$$

There is a 95% probability that the true value of the mean oxygen concentration lies within the interval of

$$5.5 \pm 2.201 * 0.65 \quad \text{mg/L}$$

$$5.5 \pm 1.4 \quad \text{mg/L}$$

$$4.1 \text{ to } 6.9 \quad \text{mg/L}$$

Notice that it is conventional to report standard errors, or values of the *t*-statistic multiplied by a standard error, with no more than two significant figures. The magnitude of the $t_{n-1;a/2}s_{\bar{x}}$ term then dictates how many significant figures should be used to report the value of the sample average. In this example, it would have been inappropriate to report the sample average with three significant figures, i.e. as 5.51 mg/L, because that implies a level of precision that is unwarranted given the uncertainty in the estimate. If the value of the standard error had been 0.065 mg/L then the $t_{n-1;\alpha/2}\ s_{\bar{x}}$ term would have been 0.14 mg/L. In this case, it would have been appropriate to report the sample average with three significant figures.

Before we conclude this section, consider that sometimes a reported confidence interval does not truly represent the uncertainty in the observation because the sample standard deviation may not capture all the possible sources of variability in the observation. To precisely estimate uncertainty in a measurement, an experimentalist must make **replicate measurements**. True replication involves redundancy in all aspects of the experiment that may contribute to error. For example, suppose an experiment is designed to infer a reaction rate by measuring the concentration of a reactant over time. The experimentalist may set up a single reactor vessel and take samples for analysis at specified points in time. To improve precision, the experimentalist may take more than one sample at a given point in time and average the measured concentrations. There may be some variation due to small inconsistencies in sampling and sample handling, but all these samples came from the same reactor. There are many more possible sources of error that have not been captured by taking replicate samples from a single reactor. The experiment itself has not been replicated. A better design would be to set up more than one reactor vessel, and ideally these would be at different time periods and different spatial locations. Sometimes the time and resources constrain the experimentalist's endeavors in replication. If this is the case, the data analyst must be cognizant of the extent to which the variation in a sample of observations represents the true uncertainty in the measurement.

Estimation of Errors in Derived Quantities

Frequently, we make experimental measurements that are used to infer the value of a quantity that is difficult to measure directly. For example, if we want to know the density of a fluid the easiest approach may be to weigh a measured volume of the fluid. The density is calculated as the weight divided by the volume. The question is how do the errors in the measurements of the weight and volume translate into error in the estimate of the density?

Consider a random variable, z, that is a function of N random variables $\{x_1, x_2, \ldots\}$. It can be shown that if we assume the errors are relatively small and there is no covariance between the variables $\{x_1, x_2, \ldots\}$ a Taylor series expansion of the error in z will produce the following expression for the variance in z (for more details, see Bevington and Robinson, 1992)

$$\sigma_z^2 = \sum_{i=1}^{N}\left(\frac{\partial z}{\partial x_i}\right)^2 \sigma_{x_i}^2 \tag{6}$$

This relationship can be used to estimate the variance in a derived quantity that is a function of independent variables provided that the sample variances of the measured variables have been estimated. Equation 6 also partitions the uncertainty in the independent variables which can provide useful information about what measurements are the most important with regard to improving the precision in the estimate of the derived quantity. Each independent variable, x_i, contributes to the variance in z in two ways, through its uncertainty i.e. σ_{x_i} , and through the mathematical sensitivity of z to x_i, i.e. the partial differential $\partial z/\partial x_i$.

Before proceeding, let's examine the concept of **covariance** in experimental data. A covariance of zero does not mean that x_1 and x_2 may not be interdependent, it simply means that the variation in one is independent of the variation in the other. For example, we know that the density of a material can be related to its weight and volume. The weight and the volume are most definitely related to each other. A larger volume must weigh more and visa versa. However, there is probably no reason to believe that the error in the measurement of the weight is in any way related to the error in the measurement of the volume.

As an example of a derived quantity that is additively related to other variables, consider the case that z is the weighted sum of two random variables, x_1 and x_2

$$z = ax_1 + bx_2 \tag{7}$$

where a and b are constants. If the covariance between x_1 and x_2 is zero, then the variance in z is

$$\sigma_z^2 = a^2\sigma_{x_1}^2 + b^2\sigma_{x_2}^2 \tag{8}$$

In general for a sum, z, that is the weighted sum of N independent variables

$$z = \sum_{i=1}^{N} a_i x_i \tag{9}$$

where the a_i's are constant coefficients for the random variables, x_i's, the variance in z is

$$\sigma_z^2 = \sum_{i=1}^{N} a_i^2\sigma_{x_i}^2 \tag{10}$$

EXAMPLE. An experiment is conducted to estimate the weight of water that has evaporated from a small pan placed in the sunlight. The experimentalist weighs out 4.0 kg of water and places it in the pan, and repeats this process four more times. After a period of time during which there has been no rain, the remaining water in the pan is weighed and found to be 16.2 kg. The estimated weight of water that has been lost due to evaporation, E, is related to amounts of water added and remaining through

$$E = 5A - R$$

where A is the weight of each individual amount of water that was added and R is the weight of water that is remaining in the pan. The experimentalist calculates the estimated value of E as 3.8 kg. Based on multiple measurements and past experience, the experimentalist estimates that the device used to measure each individual amount of water added has a standard deviation of 0.1 kg, i.e. $s_A = 0.1\ \text{kg}$. A different device was used to measure the much larger weight of the remaining water. The measurement error associated with this device is estimated to be $s_R = 0.2\ \text{kg}$. Assuming there is no covariance in the measurements of A and R, the standard deviation of the estimate of E is

$$
\begin{aligned}
s_E &= \sqrt{5^2\, s_A^2 + (-1)^2 s_R^2} \\
&= \sqrt{25(0.01) + 0.04} \\
&= \sqrt{0.25 + 0.04} \\
&= 0.54 \quad \text{kg}
\end{aligned}
$$

It makes sense that the estimated error in the derived quantity, *E*, is greater than the error in either of the variables that went into its calculation. From this calculation it is also possible to examine which experimental measurement contributes the largest source of uncertainty. In this case, the greatest uncertainty comes from the total measured weight of the water added. Despite the fact that the measuring device used for this weight measurement was more precise than that used for the measurement of *R*, the total contribution to the uncertainty is larger due to the fact that this measurement had to be made five times.

Quite often, derived quantities result from multiplicative relationships. Consider a random variable, *z*, that is the product of two random variables, x_1 and x_2, each one raised to some power

$$z = x_1^a x_2^b \tag{11}$$

One can derive an expression for the variance in *z* by differentiating equation 11 to derive the partial differentials of *z* with respect to x_1 and x_2, and then applying equation 6. Alternatively we can turn equation 11 into an additive relationship by taking the logarithm of both sides.

$$\ln z = a \ln x_1 + b \ln x_2 \tag{12}$$

Assuming x_1 and x_2 are independent (i.e. no covariance), we can apply the principles in equation 10 to the summation and get

$$\sigma_{\ln z}^2 = a^2 \sigma_{\ln x_1}^2 + b^2 \sigma_{\ln x_2}^2 \tag{13}$$

The standard deviation of the natural logarithm of a random variable is approximately equal to the **relative standard error** (also called the **coefficient of variation**), i.e.

$$\sigma_{\ln x} \approx \frac{\sigma_x}{x} \tag{14}$$

if the standard deviation for that variable is small. This approximation should look familiar to those who have studied calculus since the total differential for the logarithm of a variable is

$$d \ln x = \frac{dx}{x}$$

The approximation in equation 14 holds for relative standard errors less than 10%, i.e. $\sigma_x / x \le 0.10$, and is a good approximation even for relative standard errors up to 20%. Applying the approximation in equation 14 to equation 13 gives

$$\frac{\sigma_z^2}{z^2} = a^2 \frac{\sigma_x^2}{x^2} + b^2 \frac{\sigma_y^2}{y^2} \tag{15}$$

For multiplicative relationships, the error in a derived quantity depends not only on the errors in the independent variables and the sensitivities but also on the magnitudes of the independent variables. In general for a variable, *z*, that is the product of powers of *N* independent variables:

$$z = \prod_{i=1}^{N} x_i^{a_i} \tag{16}$$

the relative variance in *z* is

$$\frac{\sigma_z^2}{z^2} = \sum_{i=1}^{N} a_i^2 \frac{\sigma_{x_i}^2}{x_i^2} \tag{17}$$

In the event that replicate measurements of *x* have been made, the coefficient of variation, $\sigma_{\bar{X}} / \bar{x}$, is estimated by the relative standard error which is $s_{\bar{X}} / \bar{x}$.

EXAMPLE. An experiment is designed to infer the density of a liquid (ρ) from its measured mass (M) and volume (V). How do the errors in M and V translate into error in the calculated value of ρ?

$$\rho = \frac{M}{V}$$

$$\ln \rho = \ln M - \ln V$$

$$\sigma_{\ln \rho}^2 = \sigma_{\ln M}^2 + \sigma_{\ln V}^2$$

If the errors in M and V are not excessively large,

$$\frac{\sigma_\rho^2}{\rho^2} = \frac{\sigma_M^2}{M^2} + \frac{\sigma_V^2}{V^2}$$

$$\frac{\sigma_\rho}{\rho} = \sqrt{\left(\frac{\sigma_M}{M}\right)^2 + \left(\frac{\sigma_V}{V}\right)^2}$$

The relative standard error s_M / M is an estimate for σ_M / M, and likewise for V.

Suppose that the relative standard error in M is 0.01 (i.e. 1%), and the relative standard error in V is 0.05 (i.e. 5%). What is the relative standard error in ρ?

$$\frac{s_\rho}{\rho} = \sqrt{(0.01)^2 + (0.05)^2} = 0.05$$

The error in the inferred value of the density is 5%. In this case, the error in the density estimate is largely controlled by uncertainty in the measured volume. Efforts to reduce the error in the volume measurement would be the most fruitful way of improving the precision in the estimated density.

Hypothesis Testing and the t-Test

One of the most common uses of statistics is to compare a measured value with either a known value or another measured value. For example, suppose an experiment is being conducted to examine the extent to which solution chemistry affects the dissolution rate of a particular mineral. Let's assume there are two conditions of interest. There are two general ways of designing such an experiment. One possibility is to set up replicates of experimental systems under one condition and an independent set of replicate experimental systems that are observed under the other condition. The average of the measured dissolution rates from each set would be assumed to represent the dissolution rate under the respective condition. In this case, the experimentalist would be interested in comparing the two averages and inferring the extent to which the two values differ. Alternatively, a series of experimental systems could be set up all with one condition and the dissolution rates measured. Then the solution chemistry could be changed in each system and the dissolution rate measured again. The two measurements for each experimental system are clearly not independent of each other and cannot be analyzed separately. In this case, the experimentalist would examine the change in dissolution rate for each experimental system, and compute the average of the differences in the dissolution rate. This average would be compared with zero (a known value). Obviously, there are technical constraints that would favor one experimental design over the other, but both of these experimental designs would allow one to examine the effect of solution chemistry on mineral dissolution rate. It is important for the experimentalist to understand the difference in the appropriate statistical procedure for data analysis in each case.

Statistical analysis that examines differences between samples of observations is called analysis of variance (ANOVA). Analysis of variance usually refers to statistical analysis involving simultaneous comparison of multiple sets of observations, not just the comparison of two averages. When applied to comparison of two averages, or an average and a known value, the statistical procedure known as ANOVA simplifies to what is commonly called a *t*-test. For detailed discussions of ANOVA and experimental design involving multiple sets of observations, the reader is referred to Box, et al. (1978). Here we will discuss the simple case of *t*-tests for comparison of two values. In statistical *t*-tests, and in other statistical tests, the first step is to formulate the **null hypothesis**. For the case of comparison of two values, it is conventional that the null hypothesis be a statement that there is no difference between the two values. Then we analyze the data to examine the extent of the evidence to reject the null hypothesis relative to a specific alternate hypothesis.

Comparing A Sample Average with a Known Value

The example given above in which mineral dissolution rates are observed in replicate experimental systems in which the solution chemistry is changed is a case where a sample average is compared with a known value. This kind of experimental design is often called a "paired comparison". The average of the differences in the pairs of observations is compared with the value zero. Another case where one would need to compare a sample average with a known value would be, for example, if the concentration of a chemical in a solution is known with great certainty, and we are testing whether an instrument generates measurements that are consistent with the known concentration. In either case, the null hypothesis is that the true sample mean, η, is the same as the known value, η_0. Of course, the estimated mean, i.e. the sample average $\bar{x}$, will most likely be different from η_0. If there is a great deal of uncertainty in the measured sample average then there may be insufficient evidence to reject the null hypothesis.

We assume that the errors in the measured variable are IID and normally distributed. Because $\bar{x}$ has uncertainty, one can think of it as a random variable. The error in this random variable, i.e. the difference between the sample average and the true sample mean, has the *t*-distribution with n-1 degrees of freedom, scaled by the standard error of the sample average, i.e.

$$(\bar{x} - \eta) \sim t_{n-1}\, s_{\bar{x}} \tag{18}$$

where n is the number of observations used to compute the average. If the sample used to compute $\bar{x}$ is a set of paired differences, then n is the number of sets of pairs, i.e. the number of differences that are used to compute the average difference. The symbol "~" in equation 18 means "is distributed as". As is consistent with the null hypothesis, the difference in equation 18 has an expected value of zero.

To test the null hypothesis we estimate the probability associated with an **alternate hypothesis**. Possible alternative hypotheses are $\eta \neq \eta_0$, $\eta < \eta_0$, or $\eta > \eta_0$. The choice of the alternate hypothesis determines whether we conduct a one-sided *t*-test or a two-sided *t*-test. Imagine that we are interested in the alternate hypothesis that the true mean is greater than the known value, i.e. the difference $(\eta - \eta_0)$ is greater than zero. This means we want to conduct a one-sided *t*-test. The probability associated with this alternate hypothesis is:

$$\begin{aligned}
&\Pr\{\eta - \eta_0 > 0\} = \Pr\{\eta > \eta_0\} \\
&= \Pr\{(\bar{x} - \eta) \leq (\bar{x} - \eta_0)\} \\
&= \Pr\left\{\left(\frac{\bar{x} - \eta}{s_{\bar{x}}}\right) \leq \left(\frac{\bar{x} - \eta_0}{s_{\bar{x}}}\right)\right\} \\
&= \Pr\left\{t_{n-1} \leq \left(\frac{\bar{x} - \eta_0}{s_{\bar{x}}}\right)\right\} \\
&= 1 - \alpha
\end{aligned} \tag{19}$$

The term in parentheses is called the "observed" value of the *t*-statistic. It is a ratio of the observed error in the sample average (assuming the null hypothesis to be true) to the standard error, and it can be thought of as a ratio of the "signal" to "noise". If the difference between $\bar{x}$ and η_0 is large relative to the standard error in $\bar{x}$, then the probability $1-\alpha$ is large. In this case, it is very unlikely that the observed difference would occur due to random chance. One would say that there is significant evidence to reject the null hypothesis. Alternatively, we could have estimated the probability of the alternate hypothesis that the true mean is less than the known value. Both of these cases are one-sided *t*-tests. For the third alternate hypothesis, that the true mean is different from (either greater than or less than) the known value, we conduct a two-sided *t*-test. Because of the symmetry of the *t*-distribution, the probability associated with this is:

$$\Pr\left\{-\left|\frac{\bar{x} - \eta_0}{s_{\bar{x}}}\right| \leq t_{n-1} \leq \left|\frac{\bar{x} - \eta_0}{s_{\bar{x}}}\right|\right\} = 1 - \alpha \tag{20}$$

If the probability $1-\alpha$ is large then there is a great deal of evidence that the true mean is different from η_0. If the probability $1-\alpha$ is small then there is very little evidence to reject the null hypothesis and we say that there is not a statistically significant difference between η and η_0. The two-sided *t*-test is actually the most common form of the *t*-test because often one does not have *a priori* knowledge of the sign of the difference between η and η_0.

Keep in mind that there is a distinction between a difference that is statistically significant and a difference that is important. For example, one may find that there is a statistically significant difference between mineral dissolution rates at different solution chemistry conditions, but perhaps the magnitude of

this difference is too small to be important relative to other rate processes. The *t*-test can only indicate whether differences are statistically significant.

EXAMPLE. It is believed that increased partial pressure of carbon dioxide, P_{CO2}, in the atmosphere may accelerate weathering of minerals by increasing the concentration of carbonate in waters and by affecting the pH. An experiment is conducted to examine how variation in P_{CO2} affects the rate of dissolution of calcium from soil minerals. The experiment is designed by setting up a series of 3 replicate reactor vessels. The systems are identical with regard to all the factors that might affect the mineral dissolution process. Initially, the P_{CO2} is controlled at 600 Pa and the rate of Ca dissolution is determined. Then the P_{CO2} is adjusted to 1000 Pa in each of the reactor vessels and the rate of Ca dissolution is again determined. To compare the dissolution rates for each vessel, one must assume that the treatment at 600 Pa did not change the system in a significant way such that the initial conditions for the 1000 Pa treatment can be assumed to be the same as the initial conditions for the 600 Pa treatment. (Sometimes it is not possible to design experiments like this.)

The surface area-normalized dissolution rate measurements are shown in the following table. The differences between the rates for each vessel are tabulated in the last column, and the averages, sample standard deviation of the differences, and standard error of the mean differences are also shown.

	Ca Dissolution Rate [μmol m^{-2} hr^{-1}]		Difference
Reactor vessel	A; P_{CO2} = 600 Pa	B; P_{CO2} = 1000 Pa	B-A
1	347	600	253
2	96	337	241
3	174	402	228
Averages	205.7	446.3	240.7
		Sample standard deviation of the differences	12.5
		Standard error of the mean difference	7.2

What is the probability that there is a difference between the true dissolution rates at the two P_{CO2} conditions? This question calls for a two-sided *t*-test. This is appropriate if we can assume that the observed differences are IID with normally distributed errors.

To compute the probability in equation 20, we must compute the probability that the *t*-statistic with (3-1) = 2 degrees of freedom lies within the range bracketed by the positive and negative values of the observed *t*-statistic:

$$\left|\frac{\bar{x}-\eta_0}{s_{\bar{x}}}\right| = \left|\frac{240.7-0}{7.2}\right| = 33$$

In this case, the "known value" against which the average is being compared is zero. The observed value of the *t*-statistic, 33, is very large indicating a strong signal to noise ratio. One can use a table of *t*-statistics to find the probability that t_2 lies within the range of –33 to +33, but most textbooks list values of the *t*-statistic for only a few selected values of α (e.g. 0.25, 0.1. 0.05, 0.025, 0.01, 0.005). It is more useful to compute the actual

probability associated with the "observed" value. In Microsoft® Excel, for example, one can compute this probability using the TDIST function. For this experiment,

$$\Pr\{-33 \le t_2 \le 33\} = 0.999$$

In words, the probability that the true dissolution rates at the two P_{CO2} conditions are different is 99.9%. There is only a 0.1% probability that this difference would occur by random chance. These findings present compelling evidence that the mineral dissolution rate is dependent on P_{CO2}.

Comparing Two Sets of Measurements

Consider an experimental design in which mineral dissolution rates are observed in two series of replicate experimental systems where the solution chemistry for all the systems in one series is fixed to represent one condition and for all the systems in the other series is fixed to represent the other condition. This is a case where the dissolution rate measurements under the two conditions are independent of each other and the two sample averages are compared with each other. Often times in this type of experimental design, one of the experimental conditions is viewed as the "**control**". This type of experimental design is used even in the case where we think we know *a priori* how the control system will behave. For example, suppose the mineral dissolution rate of interest has been reported in the literature for a baseline condition, and the objective of our experiment is to examine the effect of changing the solution chemistry. Conceivably we could run replicate experiments only at the new solution chemistry and compare the average mineral dissolute rate with the literature value. This is usually a bad idea. A good experimentalist always runs parallel experiments, i.e. control systems, that are identical to the other experimental systems in all ways except for the treatments of interest. This eliminates the additional variation that could arise due to conducting the experiments at different times and in different labs.

Let the subscripts "A" and "B" denote the two experimental conditions, and η_A and η_B denote the true mean values of the observation variable, *x*, for the two conditions, respectively. The null hypothesis is that the difference between η_A and η_B is zero. The purpose of a *t*-test in this type of experimental design is to examine whether the difference between the observed sample averages, $\bar{x}_A$ and $\bar{x}_B$, is large relative to the uncertainty in the averages.

We assume that the errors in the measurements for the two experimental conditions are IID and normally distributed. Just as one can consider $\bar{x}$ to be a random variable, one can consider the difference $\bar{x}_A - \bar{x}_B$ to be a random variable. The error in this difference has a scaled *t*-distribution:

$$\left(\bar{x}_A - \bar{x}_B\right) - \left(\eta_A - \eta_B\right) \sim t_\nu \, s_{(\bar{x}_A - \bar{x}_B)} \tag{21}$$

where the scaling factor for the *t*-distribution is the standard error in the difference of the two sample averages. Because the null hypothesis is that the difference between η_A and η_B is zero, equation 21 can also be written

$$\left(\bar{x}_A - \bar{x}_B\right) \sim t_\nu s_{(\bar{x}_A - \bar{x}_B)} \tag{22}$$

From equation 10 we see that the variance of a difference between two variances is the sum of the variances of the variables:

$$\sigma^2_{(\bar{x}_A - \bar{x}_B)} = \sigma^2_{\bar{x}_A} + \sigma^2_{\bar{x}_B} \tag{23}$$

So the standard error in the difference is:

$$s_{(\bar{x}_A - \bar{x}_B)} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} \tag{24}$$

where s_A^2 and s_B^2 are the sample variances for the two experimental conditions. If the magnitude of the errors in the measurements of x_A and x_B are approximately equal, then the degrees of freedom associated with this standard error (ν in equation 22) is the sum of the degrees of freedom for the standard errors in each of $\bar{x}_A$ and $\bar{x}_B$, i.e. $\nu = n_A + n_B - 2$, where n_A and n_B are the numbers of values in the samples of observations from experimental conditions A and B, respectively. The common variance (also called the pooled variance) is estimated as a weighted average of the variance in x_A and x_B:

$$s^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \tag{25}$$

The standard error in the difference of the averages can then be written:

$$s_{(\bar{x}_A - \bar{x}_B)} = s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \tag{26}$$

where s is the common standard deviation computed as the square root of s^2.

To test the null hypothesis that the true means are the same, we estimate the probability associated with one of the following possible alternate hypotheses: $\eta_A \neq \eta_B$, $\eta_A < \eta_B$, or $\eta_A > \eta_B$. A two-sided *t*-test is used to test the alternate hypothesis that the true means are different from (either greater than or less than) each other. The probability associated with this is:

$$\Pr\left\{-\left|\frac{\bar{x}_A - \bar{x}_B}{s_{(\bar{x}_A - \bar{x}_B)}}\right| \leq t_\nu \leq \left|\frac{\bar{x}_A - \bar{x}_B}{s_{(\bar{x}_A - \bar{x}_B)}}\right|\right\} = 1 - \alpha \tag{27}$$

If the observed difference is large, the probability $1-\alpha$ will be large. However, in order for this difference to be statistically significant, it must be large relative to the standard error of the difference. If there is very little precision in the estimates of the averages then one still may not be able to say that that difference is statistically significant even if the observed difference is large.

> EXAMPLE. An experiment is conducted to examine how P_{CO2} affects the rate of dissolution of calcium from soil minerals. As in the previous example, there are two P_{CO2} conditions of interest, but in this case the experimental design involves two independent series of experimental systems. Three reactor vessels are operated at 600 Pa and three separate reactor vessels are operated at 1000 Pa. Imagine, for demonstration purposes, that the mineral dissolution rates that are observed in these six systems are the same values as those in the previous example.

	Ca Dissolution Rate [μmol m^{-2} hr^{-1}]	
	A; P_{CO2} = 600 Pa	B; P_{CO2} = 1000 Pa
	347	600
	96	337
	174	402
Average	205.7	446.3
Sample standard deviation	128.5	137.0
Sample variance	16502	18766

The magnitude of the observed difference in the averages is:

$$\left|\bar{x}_A - \bar{x}_B\right| = 240.7\ \mu\text{mol m}^{-2}\ \text{hr}^{-1}$$

What is the probability that there is a difference between the true mean dissolution rates at the two P_{CO2} conditions? As with the previous example, a two-sided *t*-test is appropriate. This is appropriate with an assumption that the errors in the observed mineral dissolution rates are IID and normally distributed.

The observed *t*-statistic in equation 27 requires the standard error of the difference in the averages. Because we have assumed the underlying measurement errors for conditions A and B are "identically distributed", this implies that the variances are equal. The estimated common variance and common standard deviation are:

$$s^2 = \frac{(3-1)*16502 + (3-1)*18766}{3+3-2} = 17634$$

$$s = \sqrt{17634} = 132.8\ \mu\text{mol m}^{-2}\ \text{hr}^{-1}$$

The standard error in the difference of the averages is:

$$s_{(\bar{x}_A - \bar{x}_B)} = 132.8\sqrt{\frac{1}{3}+\frac{1}{3}} = 108.4\ \mu\text{mol m}^{-2}\ \text{hr}^{-1}$$

The degrees of freedom associated with this standard error is $\nu = 3+3-2 = 4$. With a null hypothesis that the true difference between the means is zero, the observed *t*-statistic is:

$$\left|\frac{\bar{x}_A - \bar{x}_B}{s_{(\bar{x}_A - \bar{x}_B)}}\right| = \frac{240.7}{108.4} = 2.22$$

The probability of the alternate hypothesis is estimated as:

$$\Pr\left\{-2.22 \le t_4 \le 2.22\right\} = 0.91$$

In words, the probability that the true dissolution rates at the two P_{CO2} conditions are different is 91%. There is a 9% probability that this difference would occur by random chance.

Comparing the Two Types of Experimental Designs

Before concluding this section, let's compare the two types of experimental designs described above. Two sources of variation determine the observation of mineral dissolution rates. First there is the variation that is caused by the change in solution chemistry. Second there is the variation caused by random error that produces different observations for replicate experiments. A good experimental design is one that maximizes the variation due to change in solution chemistry (the "signal") relative to the variation due to random error (the "noise"). If the experiment is conducted in a way in which each replicate experimental system is observed first at one condition and then at the other condition then the variation across replicate experimental systems is **blocked** from the observed variable. Imagine that one of the reactor vessels is closer to the window where the temperature is slightly colder than the average room temperature. The mineral dissolution rates that are measured for that vessel may be biased due to this temperature difference relative to the other vessels, but this will be true for the mineral dissolution rates at both conditions. The difference in mineral dissolution rates is likely much less sensitive to temperature than the actual rates. This is demonstrated in the example of the paired comparisons. The variation in the observations for a given P_{CO2} condition is fairly large. Notice that the ranges of the observations for the two P_{CO2} conditions overlap. However, there is much less variation in the differences in the dissolution rates. This experimental design reduces the effect of variation across replicate experimental systems on the inferred difference between the two treatments. In the paired comparison, the difference between the means is much more statistically significant than in the case of the comparison between two averages from independent series of experiments. In general, given a choice between these two types of experimental designs, paired comparisons should be used if other technical constraints don't preclude this design. For additional discussion on the advantages of paired comparisons, see Berthouex and Brown (1994).

Linear Regression Analysis

Statistical Regression in Mathematical Modeling

We often wish to use experimental data to develop a mathematical model between two or more variables. We can distinguish between two applications of statistics for mathematical modeling. First, if we have a variable y that we believe to be related to a variable x then we could make a series of n measurements of the pair $\{x_i, y_i\}$ where i is an index that runs from 1 to n. If we have no *a priori* knowledge of the possible mathematical relationship between y and x, then we can use statistical techniques to estimate a mathematical relationship that captures the variation of y with x. Such a mathematical model is called a **statistical model**, which is purely empirical. An example of a useful empirical model is a power-series polynomial.

$$y = a_1 + a_2 x + a_3 x^2 + \cdots + a_m x^{m-1} \tag{28}$$

where $\{a_1 \ldots a_m\}$ are the parameters of the function.

It is more common that a mathematical relationship between the variables is known (or postulated) based on theoretical principles. For example, based on past experience and the principles of chemical thermodynamics, we know that in a system containing water in equilibrium with an organic liquid phase (e.g. octanol), for dilute concentrations the concentration of a chemical in the water phase, y, is linearly related to the concentration of that chemical in the organic liquid phase, x

$$y = ax \tag{29}$$

In this case, a is the model parameter describing the proportionality between the two concentrations. Another example of a theoretical model is the Freundlich isotherm which is often used to describe chemical equilibrium in a system of water in contact with a sorbing medium such as soil or sediment. The Freundlich model relates the concentration of a chemical in the water phase, y, to the concentration of that chemical in the sorbing medium, x, according to the following function

$$y = a_1 x^{a_2} \tag{30}$$

This model has two parameters, a_1 and a_2.

Both of these applications of statistics are examples of **regression**, in which a mathematical model has been established (either with a theoretical basis or not) and statistical regression is used to estimate the parameters of the model based on a set of data. The examples given above are for the case where there is one dependent variable y and one independent variable x. The mathematical modeling problem may involve multiple dependent variables and/or multiple independent variables. The case of multiple dependent variables can be a difficult problem and involve complex regression techniques. The case of multiple independent variables is a simpler problem and often can be handled with little more complexity than the case of a single independent variable. In general, the mathematical model we consider here is

$$y = f\left(x_1, x_2, \ldots, x_l; a_1, a_2, \ldots, a_m\right) \tag{31}$$

where y is the single dependent variable, x_1 through x_l are the l independent variables, and a_1 through a_m are the m model parameters. The general regression problem is this: given a mathematical model and a set of data, i.e. a series of n observations of the set $\left\{x_{1i}, x_{2i}, \ldots, x_{li}, y_i\right\}$, what are the values of the parameters $\left\{a_1, a_2, \ldots, a_m\right\}$ that produces a model that best describes the relationship between the dependent and independent variables. Statistical regression is also referred to as **model calibration**, data fitting, data reduction, or parameter estimation. Sometimes the phrase "curve fitting" is used specifically in reference to statistical regression of data to a polynomial function.

Linear Least Squares Regression

In the previous section we defined statistical regression as an activity that produces a model that "best describes the relationship" between y and the independent variables. Conventionally, the criterion we use to define "best" derives from the **method of maximum likelihood**. In words, the method of maximum likelihood seeks the values of the parameters that maximize the probability of having obtained the set of observations in hand. If we assume that the errors in y are independent and normally distributed, the method of maximum likelihood leads to a goodness-of-fit parameter, X^2 ("chi-squared"), which is defined as

$$X^2 \equiv \sum_{i=1}^{n} \left(\frac{y_i - f\left(x_{1i}, x_{2i}, \ldots, x_{li}; a_1, a_2, \ldots, a_m\right)}{\sigma_i} \right)^2 \tag{32}$$

where σ_i is the standard deviation associated with the ith value of y. The "best" values of the parameters are those that minimize X^2. It is quite often the case that we do not know σ_i. If we make a further assumption that the observations of y are identically distributed, i.e. not only do all values have the same underlying error structure (in this case a normal distribution) but they also have the same magnitude of error, then the value of σ_i is the same for all i. Taken together, the three assumptions are stated concisely as the errors are IID and normally distributed. With this assumption, X^2 becomes

$$X^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - f\left(x_{1i}, x_{2i}, \ldots, x_{li}; a_1, a_2, \ldots, a_m\right)\right)^2 \tag{33}$$

where σ is the common standard deviation in y. Minimizing X^2 is now a matter of finding the values of $\{a_1, a_2, \ldots, a_m\}$ that minimize the summation on the right hand side of equation 33. This form of statistical regression is called **least squares regression**.

Experimentalists should be aware of the times when it is not appropriate to assume that all the measurements have the same standard deviation. If one is using an instrument for which the error is proportional to the magnitude of the measurement then the errors will be small for measurements of smaller values and vice versa. Similarly, if the data to be used in the regression were obtained via different measurement techniques then the errors can be substantially different. For example, if low values of aqueous concentration are measured using a calibration curve designed specifically for a low concentration range, and high values of aqueous concentration are measured using a calibration curve designed for a high concentration range, then it is quite likely that they two sets of observations will have different measurement errors. The errors may be constant in a relative sense (e.g. the % error), but the absolute errors may be very different. After the regression has been performed, one can test the assumption of common errors by examining a plot of the residuals (i.e. estimated errors) versus the x variable. The residuals should have a random appearance and not appear to be related to the value of x. If the experimentalist has good reason to be concerned about the lack of a common error in the observations of y then the extra effort should be taken to estimate the errors in the observations. In this case, regression should be performed by minimizing the X^2 function in equation 32. This is a form of **weighted least squares regression**.

If the function $f\left(x_{1i}, x_{2i}, \ldots, x_{li}; a_1, a_2, \ldots, a_m\right)$ is linear in the parameters then the summation in equation 33 can be analytically differentiated with respect to each of the parameters. Setting these differentials equal to zero produces m linear equations which can be solved simultaneously for the values of the parameters that minimize X^2. This form of statistical regression is called **linear least squares regression**. Note that the model function need not be linear in the independent variables in order to use linear least squares regression. The polynomial function shown in equation 28 is nonlinear in x but it is linear in the parameters. Consequently this function can be easily differentiated to come up with analytical expressions for the optimum values of the parameters. An example of a function that is nonlinear in the parameters is the Freundlich isotherm (equation 30). In its current form, this model cannot be fit using linear least squares regression. A transformation of the Freundlich equation generates a form that is linear in the parameters, although nonlinear in the variables:

$$\log y = \log a_1 + a_2 \log x \tag{34}$$

The parameters of this equation, i.e. $\log a_1$ and a_2, can be estimated using linear least squares regression by regression to the transformed data $\{\log x_i, \log y_i\}$. However, caution is advised when linearizing models for the purpose of conducting linear least squares regression. The assumption that must be satisfied in this case is that the errors in the transformed variables are IID and normally distributed. If the errors in the untransformed variables $\{x_i, y_i\}$ are IID and normally distributed then the errors in the transformed variables most certainly are not. If this is the case, then such models should be calibrated using nonlinear regression techniques.

Consider the simplest case where y is linearly related to a single independent variable x,

$$y = a_1 + a_2 x \tag{35}$$

Because this function is linear in a_1 and a_2, linear least squares regression can be applied. (The function also happens to be linear in x, but this is not a necessary condition for linear least squares regression.) The X^2 is

$$X^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - a_1 - a_2 x_i\right)^2 \tag{36}$$

Differentiating equation 36 with respect to a_1 and a_2 and setting these differentials equal to zero produces the following equations for the optimum values of the parameters:

$$a_1 = \bar{y} - a_2\bar{x} \tag{37}$$

$$a_2 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \tag{38}$$

where the summations are taken from i=1 to n. Most experimentalists need not use equations 37 and 38 directly as there are numerous mathematical and statistical software packages that have these functions built-in. For example, in Microsoft® Excel these functions can be applied to a sample of data using the "Regression" module in the "Data Analysis" tool which can be used for single-variable or multi-variable linear least squares regression.

One of the assumptions that is implied in the application of the maximum likelihood approach to statistical regression is that there is error only in the y variable and there is no error in the x variable(s). In experimentation it is quite often the case that both x and y are measured variables and thus both are subject to error. The x variable may be something that has negligible error, like a measurement of time or length that can be recorded precisely. If there is appreciable error in the estimate of the x values, part of the uncertainty in the y value is due to the fluctuation of x and its effect on the value of y. Thus the error in the estimate of y at a given x value is the sum of the measurement error for y plus the error propagated by the uncertainty in x. We can still apply least squares regression if the portion of the uncertainty in y is small relative to its own measurement error. This is equivalent to saying that the x value is known to much greater precision than the y value. One should always perform statistical regression by assigning the variable with the greatest measurement error as the dependent variable, or the "regression variable", even if it means writing the mathematical function in a form that is inverse to its conventional presentation.

EXAMPLE. An experiment is conducted to determine the Freundlich isotherm parameters for sorption of tetrachloroethylene onto a particular soil sample. Written using conventional notation, the Freundlich equation is

$$q = K_F C^n$$

where the variables are q, the mass of solute sorbed per unit mass of soil, and C, the solute concentration in the aqueous phase. The parameters are K_F, the capacity coefficient and n, the exponent constant. Assume that log-transformed values of the variables have errors that are IID and normally distributed. This allows linear least squares regression using the linearized model:

$$\log q = \log K_F + n\log C$$

A series of ten experimental systems are set up which are identical in all respects except for the mass of tetrachloroethylene added. After equilibration under constant temperature conditions, the solute concentration in the aqueous phase is measured and the mass of solute associated with the solid phase is determined. The measurement error in q is

believed to be much larger than the measurement error in C so it is appropriate that ln q is the dependent variable, i.e. the "regression variable". The data are shown in the table below.

i	C [μg/L]	q [μg/g]	log C	log q
1	9.99	0.512	1.00	-0.291
2	21.6	0.435	1.34	-0.362
3	44.9	2.29	1.65	0.359
4	102.3	2.37	2.01	0.375
5	216.1	5.04	2.33	0.702
6	463.5	11.6	2.67	1.065
7	986.0	19.9	2.99	1.298
8	2130	22.4	3.33	1.350
9	4538	140.0	3.66	2.146
10	9852	122.9	3.99	2.090

The "Regression" tool in Microsoft® Excel was used to estimate the model parameters. A portion of the output is shown below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9809162
R Square	0.962196592
Adjusted R Square	0.957471166
Standard Error	0.181425915
Observations	10

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	6.702263327	6.702263327	203.6211
Residual	8	0.2633229	0.032915363	
Total	9	6.965586228		

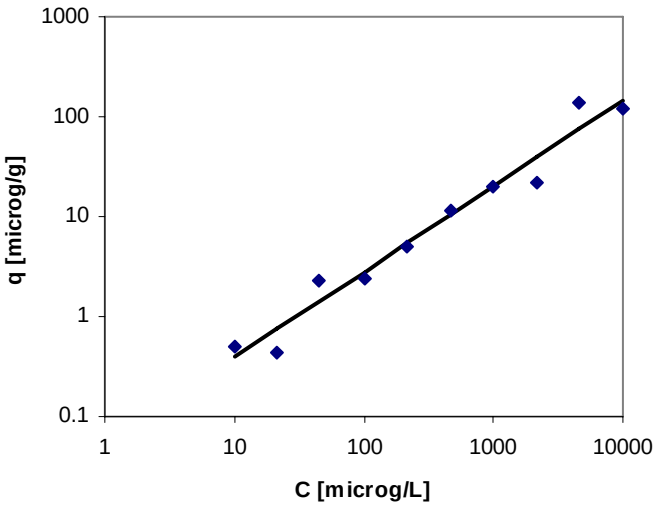
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-1.26712979	0.160599963	-7.8899756	4.82E-05
X Variable 1	0.857214638	0.060072839	14.26958761	5.67E-07

From the column labeled "Coefficients", we find the estimated parameter values:

$$\log K_F = -1.27 \qquad K_F = 0.054 \text{ L/g}$$

$$n = 0.86$$

The following plot shows the measurements of q and C plotted in logarithmic space along with the regression line.



The plot above shows no error bars on the data points. That is because the error in an individual measurement of log q is unknown. In order to estimate the error for a given value of q, one would need to set up replicate experimental systems with the same total mass of tetrachloroethylene. This kind of replication is rarely done when data are collected for regression purposes. It is possible that to determine the measured value of q, the experimentalist might take multiple samples from each experimental system or analyze a given sample more than once. These actions improve precision of the measurement but the variance in these measurement do not represent the true variance. Sometimes error bars that are shown in plots of experimental data are based on sample replications or only the analytical instrument. It is OK to put such error bars on a plot of data provided that it is explicitly stated that these error bars likely under represent the true error. The true error can be determined only by replication of the experiment.

Regression analysis can be used to estimate the error in an individual measurement of the dependent variable. The variance in the y variable that has not been explained by the regression model is the **residual sum of squares**, S_R:

$$S_R = \sum_{i=1}^{n} \left(y_i - f\left(x_{1i}, \ldots, x_{li}; a_1, \ldots, a_m\right)\right)^2 \tag{39}$$

From equation 33, we see that this is the quantity that is minimized in the regression so the parameters in equation 39 are the best-fit values. For a two-parameter linear model as in equation 35,

$$S_R = \sum_{i=1}^{n} \left(y_i - a_1 - a_2 x_i \right)^2 \tag{40}$$

In Microsoft® Excel regression output, this quantity is found in "SS" column in the row labeled "Residual". The degrees of freedom associated with this summation is the number of observations minus the number of parameters, n-m. In Microsoft® Excel regression output, this quantity is found in the "df" column in the row labeled "Residual". Dividing the residual sum of squares by the residual degrees of freedom gives a quantity that has an expected value of σ^2, the common variance of measurements of the dependent variable. This is predicated on an assumption that the parameterized model is an accurate representation of the true underlying relationship between y and x. This quantity, s^2, is called the **residual mean square** (also called "mean squared error"):

$$s^2 = \frac{S_R}{n-m} \tag{41}$$

In Microsoft® Excel regression output, this quantity is in the "MS" column in the row labeled "residual". The square root of s^2 gives an estimate of the common standard deviation of measurements of y.

EXAMPLE. For the regression example above, the residual sum of squares and residual mean square are obtained from the regression output

$$S_R = 0.263$$

$$s^2 = \frac{0.263}{8} = 0.033$$

Assuming the calibrated model is an accurate representation of the relationship between log q and log C, the square root of the residual mean square is an estimate of the standard deviation in measurements of log q.

$$s = \sqrt{0.033} = 0.18$$

Keep in mind that the assumption made about the measurements in log C is that they are error-free.

If the value of s^2 is known *a priori* or it is estimated from replicate experiments for fixed values of the x variable, then the residual mean square can be compared to this value to examine the extent to which the model fits the data. The statistical test to compare these two variances is an F-test, and in this context it is called a **lack-of-fit test**, or a "goodness-of-fit test". Without an independent estimate of s^2 one can say nothing about whether the data fit the model, although this is a commonly made misinterpretation of the regression statistics.

An alternative, and very popular, means of quantifying the extent to which the variation in the data have been captured by the model is to compute the R-square and the adjusted R-square. The R-square is the ratio of the variance in y measurements that is explained by the fitted model, i.e. the regression sum of squares (S_M), to the overall variance of the y measurements, i.e. the total corrected sum of squares (S_D). Both quantities are found in the Microsoft® Excel regression output in the "SS" column. Because the total corrected sum of squares is the sum of the S_M and S_R, the R-square can be written in terms of S_R and S_D.

$$R^2 = \frac{S_M}{S_D} = 1 - \frac{S_R}{S_D} \tag{42}$$

When R^2 is close to unity then we say that a large portion of the overall variation in the data are explained by the model. It means that y is related to x, or highly correlated to x. It does not mean that the model that has been specified is the correct model. The only way to examine if a model is the correct representation of the underlying functional relationship is with an independent estimate of σ^2 and a lack-of-fit test.

The adjusted R-square takes into account the degrees of freedom associated with each sum of squares:

$$R^2_{\text{ADJ}} = 1 - \frac{S_R/(n-m)}{S_D/(n-1)} = 1 - \frac{s^2}{S_D/(n-1)} \tag{43}$$

This term is often reported when doing statistical regression using polynomial models. When the modeler has a choice of the number of parameters in the model, then it is possible to improve the regression

merely by increasing the number of parameters. Often this activity results in meaningless models that "overfit" the data. The idea behind the adjusted *R*-square is that it has a penalty term for the number of parameters. If *m* is large then the residual degrees of freedom *n-m* may be small. The adjusted *R*-square will reflect the fact that the a large portion of the overall variance has been explained only because of the large number of parameters. The *R*-square will not reflect this.

EXAMPLE. For the regression example above, the *R*-square can be computed from the values in the "SS" column, or it can be read directly from the "Regression Statistics":

$$R^2 = \frac{6.70}{6.97} = 1 - \frac{0.26}{6.97} = 0.96$$

Typically *R*-square values greater than 0.90 are considered indicators of a high degree of correlation between the variables. This *R*-square of 0.96 indicates that log *q* and log *C* are highly correlated. Note that the adjusted *R*-square (also in the "Regression Statistics") is very close to the *R*-square. For this model the number of parameters is small, and the residual degrees of freedom, 8, is almost equal to the total degrees of freedom, 9.

Estimation of the Errors in the Parameters

Because the estimated parameters in linear least squares regression are linear functions of normally distributed variables (equations 37 and 38), both a_1 and a_2 in equation 35 can be considered normally distributed random variables. As such, for each parameter, the error in the parameter divided by its standard error is a random variable that is distributed according to the *t*-distribution. The formulas for computing the standard errors in the fitted parameters are given in most textbooks. These quantities can be found in the Microsoft® Excel regression output in the "Standard Error" column next to the estimated parameter values. The degrees of freedom associated with both standard errors is *n-m*. The "t Stat" is the "observed" value of the *t*-statistic for each parameter, according to the null hypothesis that the true value of the parameter is zero. The standard errors of the parameters can also be used with the *t*-distribution to compute confidence intervals for the parameters. These are included in the Microsoft® Excel regression output in the column "Lower 95%" and "Upper 95%" (not shown here).

EXAMPLE. For the regression example above, the observed value of the *t*-statistic for the log K_F parameter is

$$\left| \frac{-1.27 - 0}{0.161} \right| = 7.9$$

The probability associated with the alternate hypothesis that the true value of log K_F is different from zero is determined from a two-sided *t*-test:

$$\Pr\{-7.9 \le t_8 \le 7.9\} = 0.99995$$

The difference between this value and unity, i.e. α, can be found in the Microsoft Excel regression output under the "P-value" column. Clearly, the estimated value of log K_F is statistically very different from zero.

BIBLIOGRAPHY

Berthouex, P. M.; L. C. Brown. *Statistics for Environmental Engineers*. Lewis Publishers, 1994.

Bevington, P. R.; D. K. Robinson. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, Inc. 1992.

Box, G. E. P.; W. G. Hunter; J. S. Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. John Wiley & Sons, 1978.

Devore, J.; N. Farnum. *Applied Statistics for Engineers and Scientists*. Duxbury Press, 1999.

McBean, E. A.; F. A. Rovers. *Statistical Procedures for Analysis of Environmental Monitoring Data & Risk Assessment*. Prentice Hall, 1998.