Teaching What Degrees of Freedom Are In Statistics

Kenneth Sutrick, Murray State University, Murray, Kentucky, USA

ABSTRACT

One of the most confusing topics in statistics is degrees of freedom. Everyone is taught that the sample standard deviation has $n-1$ degrees of freedom. Why is this the case since you use all n data points to compute the standard deviation? The paper shows why this is the case by showing that the sample standard deviation can be broken down into $n-1$ independent parts and that the last deviation can be absorbed into these independent parts. After that a related technique shows what degrees of freedom are about in the important cases of regression and analysis of variance.

Keywords: sample standard deviation, degrees of freedom, degrees of freedom in multiple regression, degrees of freedom in analysis of variance.

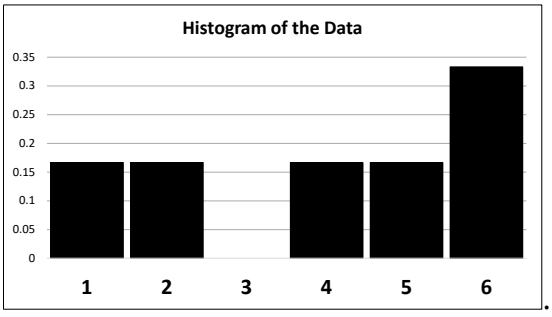
INTRODUCTION

In introductory statistics degrees of freedom are a mystery. In beginning statistics degrees of freedom are first encountered in the sample standard deviation which is used in t-confidence intervals and t-tests for the population mean μ. Students are taught that the sample standard deviation, in this case, causes the t-distribution to have $n-1$ degrees of freedom when there are n data points. The sample standard deviation is also used to find confidence intervals for the population standard deviation σ where this confidence interval is calculated using the chi-square distribution which also has degrees of freedom. The next place where degrees of freedom are usually met is in regression where they are even more complicated. This paper presents ways of teaching degrees of freedom that can be understood in introductory classes. It is possible to understand what degrees of freedom are in an elementary statistics context, for the most part, without having to go all the way to advanced statistics and mathematics. The first section of the paper discusses degrees of freedom in the standard deviation context. The second section talks about degrees of freedom in regression, and the third section covers degrees of freedom for the *ANOVA* Table in regression. The fourth section covers degrees of freedom in Analysis of Variance problems.

DEGREES OF FREEDOM FOR THE STANDARD DEVIATION

Suppose you have a data set with $n = 6$ data points: $X_1 = 5$, $X_2 = 1$, $X_3 = 4$, $X_4 = 2$, $X_5 = 6$, and $X_6 = 6$. These data points came from a population that is being studied and give general information about the population. For example you can use this data to draw a histogram, as in Figure 1, to infer the shape of the population.

Figure 1: The Data Histogram



This histogram is skewed to the left suggesting that the population may be skewed to the left. The X-values in this section will be considered as variables since they depend on what data points from the population have been collected.

It is almost always the case that the original data is transformed to get other information about the population such as its center and such as how spread out the population is. The measures from the data of center and spread, the sample average and the sample standard deviation, are used to estimate the unknown μ and σ of the population that is being studied. These measures are calculated below for the above data.

The sample average is:

$$SampleAverage = \bar{X} = \frac{5+1+4+2+6+6}{6} = \frac{24}{6} = 4.$$

The spread in the data is measured around the average. The first step in computing the standard deviation is to find how far each data point is from the average. The rest of the steps, as are taught in statistics, are given in Table 1.

Table 1: Computing the Standard Deviation

Data X	Deviations from the Average $X-\bar{X}$	Squared Deviations $(X-\bar{X})^2$
5	$d_1 = 5-4 = 1$	${d_1}^2 = 1$
1	$d_2 = 1-4 = -3$	${d_2}^2 = 9$
4	$d_3 = 4-4 = 0$	${d_3}^2 = 0$
2	$d_4 = 2-4 = -2$	${d_4}^2 = 4$
6	$d_5 = 6-4 = 2$	${d_5}^2 = 4$
6	$d_6 = 6-4 = 2$	${d_6}^2 = 4$
	Sum = 0	Sum = *SSX* = 22

(SSX = Sum of Squares X).

The sample standard deviation is defined as:

$$SampleStandardDeviation = s = \sqrt{\frac{SSX}{n-1}} = \sqrt{\frac{{d_1}^2+{d_2}^2+{d_3}^2+{d_4}^2+{d_5}^2+{d_6}^2}{6-1}} = \sqrt{\frac{22}{6-1}} \approx 2.1.$$

From these calculations one would estimate that the center of the population is about 4 and that the spread in the population around 4 is about 2.1. This section of the paper explains why here in the calculation of "s" you should divide by $6-1$ and not 6, or in the general case what the difference between n and $n-1$ is all about.

The standard deviation is interpreted as follows: *The standard deviation represents a typical distance or deviation of a data point from the average.* The second column in Table 1, containing the deviations from the average, $X-\bar{X}$, tells you how far each data point is from the average. For example the data point 5 with a deviation of 1 is one unit from the average of 4, while the data point 1 with a negative deviation is three units below the average. A typical number in the column of deviations is 2 (ignoring the minus signs). The fact that standard deviation of about 2.1 is close to this typical number 2 is not a coincidence. This always happens when you compute the standard deviation. It will always be the case that the standard deviation will be close to a typical number on the list of $X-\bar{X}$, and this is a useful way of understanding what the standard deviation measures.

The reason the standard deviation is calculated as in Table 1, with the subtracting the average, then squaring, adding, and everything else, has to do with the normal curve. If your data is normal you can prove that the optimal way to measure spread is via the above calculation. In this data the quantity SSX in the computation of s is calculated from the six deviations: d_1, d_2, …, d_6 yet is said to have 5 degrees of freedom. To start to see why this is the case, in Table 1 you can see that the deviations, defined as: $d_1 = X_1 - \bar{X}$, $d_2 = X_2 - \bar{X}$, …, $d_6 = X_6 - \bar{X}$, satisfy $d_1 + d_2 + d_3 + d_4 + d_5 + d_6 = 0$. The consequence of this is that the deviations are related at least a little, so that in this case $d_6 = -d_1 - d_2 - d_3 - d_4 - d_5$. This is evident in Table 1 since: $d_6 = 2 = -1-(-3)-0-(-2)-2$.

Therefore you do not even need d_6 since you can always get it from the other five deviations if necessary. The sample standard deviation here is really a function of the five numbers: d_1, d_2, …, d_5,

$$s=\sqrt{\frac{SSX}{n-1}}=\sqrt{\frac{(1)^2+(-3)^2+(0)^2+(-2)^2+(2)^2+(2)^2}{6-1}}=\sqrt{\frac{{d_1}^2+{d_2}^2+{d_3}^2+{d_4}^2+{d_5}^2+(-d_1-d_2-d_3-d_4-d_5)^2}{6-1}}.$$

This is the first indication that s here has 5 degrees of freedom and not 6, the number of squares in Table 1. [The fact that the deviations add to zero is true for every data set since once you take away the middle what's left adds to zero.]

A good way to think about degrees of freedom is to think of a degree of freedom as a piece of information. The six data points: X_1, X_2, X_3, X_4, X_5, and X_6 have 6 pieces of general information about the population or 6 degrees of freedom. Given the original data, X_1, X_2, X_3, X_4, X_5, X_6, you can calculate $\bar{X}$, d_1, d_2, d_3, d_4, and d_5. However since $X_1=\bar{X}+d_1$, $X_2=\bar{X}+d_2$, …, $X_6=\bar{X}+d_6=\bar{X}-d_1-d_2-d_3-d_4-d_5$, then given $\bar{X}$, d_1, d_2, d_3, d_4, and d_5 you can get back X_1, X_2, X_3, X_4, X_5, and X_6. Since you can go back and forth between both sets (of six numbers), they must contain the same amount of information about the population. Consequently the numbers: $\bar{X}$, d_1, d_2, d_3, d_4, and d_5 must contain 6 pieces of information, it is just that the information is in a different form. Instead of general information about the population, the numerical value of $\bar{X}$ has one degree of freedom that measures the center of the population, the values of d_1, d_2, d_3, d_4, and d_5 have 5 degrees of freedom that measure the spread in the population. The original data could be transformed in other ways to get other information about the population such as its skewness and kurtosis and you could even divert some of the information in the d's toward that purpose. However if there are six data points there is a maximum of six independent pieces of information available to study the population. The researcher can then decide to what purposes that information should be used. For studying the degrees of freedom in the standard deviation we will assume that all information other than that in $\bar{X}$ will be used for measuring spread. In statistics it is common to assume that the data points X_1, X_2, … are statistically independent so that each point gives additional information about the population. Under this assumption you can prove that d_i and d_j are correlated, this is evident since the d's sum to zero. This basically means that some of the information about spread that is in d_i is also contained in d_j. It is this fact that makes studying the behavior of SSX more complicated. In the next paragraph SSX is separated into its uncorrelated parts. Before moving on, a key idea to notice in this paragraph is that to measure center and spread you must transform the original data. For example, you can write: $\bar{X}=(1/6)X_1+\cdots+(1/6)X_6$, $d_1=(5/6)X_1-(1/6)X_2-\cdots-(1/6)X_6$, and so on.

The sample average, sample standard deviation, sample skewness and kurtosis are all transformations of the data. Transformations are a constant feature of statistics. The idea that the sample standard deviation above has 5 degrees of freedom can be made even more exact if you consider a slightly different and perhaps more complicated transformation of the original data. We will define the transformation first and then discuss the properties of the transformation. Let

$$\begin{aligned} Z_1&=\bar{X},\\ Z_2&=\frac{X_1-X_2}{\sqrt{2}},\\ Z_3&=\frac{X_1+X_2-2X_3}{\sqrt{6}},\\ Z_4&=\frac{X_1+X_2+X_3-3X_4}{\sqrt{12}},\\ Z_5&=\frac{X_1+X_2+X_3+X_4-4X_5}{\sqrt{20}},\\ Z_6&=\frac{X_1+X_2+X_3+X_4+X_5-5X_6}{\sqrt{30}}. \end{aligned} \qquad (1)$$

For understanding degrees of freedom it is not necessary to know why the Z's are defined this way or even how these Z's were calculated, they come from the subject of linear algebra. The details of this can be left to advanced statistics and advanced mathematics. For understanding degrees of freedom what you need to know is that the Z's defined this way have the properties detailed in the next two paragraphs and that it is always possible to find such Z's for every data set. These Z's are uncorrelated. A little about where the Z's come from is discussed below in the general case, but the square roots turn out to be absolutely necessary for the above and for everything below.

In Equation 1 from the X's you compute the Z's. However it is also true that given the Z's you can get the X's back. For example: $X_6 = Z_1 - \sqrt{\frac{5}{6}}Z_6,\ X_5 = Z_1 - \sqrt{\frac{4}{5}}Z_5 + \sqrt{\frac{1}{30}}Z_6$ and it is possible to write down the formulas for X_4, X_3, X_2, and X_1 if you need to. The exact formulas for the X's in terms of the Z's are not important. The important fact here is that you can go from the the X's to the Z's and vice versa, so accordingly they must contain the same amount of information about the population, just in different forms. Even more specifically, given d_1, …, d_6 you can find Z_2, …, Z_n and vice versa. For example: $d_1 = \frac{1}{\sqrt{2}}Z_2 + \frac{1}{\sqrt{6}}Z_3 + \frac{1}{\sqrt{12}}Z_4 + \frac{1}{\sqrt{20}}Z_5 + \frac{1}{\sqrt{30}}Z_6,\ d_2 = -\frac{1}{\sqrt{2}}Z_2 + \frac{1}{\sqrt{6}}Z_3 + \frac{1}{\sqrt{12}}Z_4 + \frac{1}{\sqrt{20}}Z_5 + \frac{1}{\sqrt{30}}Z_6$, and so on and $Z_2 = \frac{1}{\sqrt{2}}d_1 - \frac{1}{\sqrt{2}}d_2 + 0d_3 + 0d_4 + 0d_5 + 0d_6,\ Z_3 = \frac{1}{\sqrt{6}}d_1 + \frac{1}{\sqrt{6}}d_2 - \frac{2}{\sqrt{6}}d_3 + 0d_4 + 0d_5 + 0d_6$ etc. Therefore d_1, …, d_6 and Z_2, …, Z_n contain the same information, in this case all of this information is about spread. The important property of the Z's is that it is always the case that no matter what the X's are, the following equation holds:

$$SSX = {d_1}^2 + {d_2}^2 + {d_3}^2 + {d_4}^2 + {d_5}^2 + {d_6}^2 = Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2. \quad (2)$$

So that

$$s = \sqrt{\frac{Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2}{6-1}}.$$

The information in the 6 correlated d's has been transferred to the last 5 uncorrelated Z's. The Z's have extracted the independent pieces of information about spread that are available in the d's.

To check Equation 2, if $X_1 = 5, X_2 = 1,\ X_3 = 4,\ X_4 = 2,\ X_5 = 6$, and $X_6 = 6$, then $Z_1 = 4, Z_2 = \frac{4}{\sqrt{2}}, Z_3 = \frac{-2}{\sqrt{6}}, Z_4 = \frac{4}{\sqrt{12}}, Z_5 = \frac{-12}{\sqrt{20}}$, and $Z_6 = \frac{-12}{\sqrt{30}}$, and it is easy to verify that: $SSX = {d_1}^2 + {d_2}^2 + {d_3}^2 + {d_4}^2 + {d_5}^2 + {d_6}^2 = 12+-32+02+-22+22+22=22=162+46+1612+14420+14430=Z22+Z32+Z42+Z52+Z62.$ Just to check that you can get the d's from the Z's, using the formulas in the previous paragraph: $d_1 = \frac{1}{\sqrt{2}}\left(\frac{4}{\sqrt{2}}\right) + \frac{1}{\sqrt{6}}\left(\frac{-2}{\sqrt{6}}\right) + \frac{1}{\sqrt{12}}\left(\frac{4}{\sqrt{12}}\right) + \frac{1}{\sqrt{20}}\left(\frac{-12}{\sqrt{20}}\right) + \frac{1}{\sqrt{30}}\left(\frac{-12}{\sqrt{30}}\right) = 1$, and so on. [If the data were different, for example if: $X_1 = 5, X_2 = 1, X_3 = 3,\ X_4 = 2,\ X_5 = 6$, and $X_6 = 7$, then $\bar{X} = 4, d_1 = 1, d_2 = -3, d_3 = -1, d_4 = -2, d_5 = 2, d_6 = 3, Z_1 = 4, Z_2 = \frac{4}{\sqrt{2}}, Z_3 = 0, Z_4 = \frac{3}{\sqrt{12}}, Z_5 = \frac{-13}{\sqrt{20}}$, and $Z_6 = \frac{-18}{\sqrt{30}}$. Again you can check that: $SSX = (1)^2 + (-3)^2 + (-1)^2 + (-2)^2 + (2)^2 + (3)^2 = 28 = \frac{16}{2} + 0 + \frac{16}{12} + \frac{169}{20} + \frac{324}{30} = Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2$. Again to check that you can get the d's from the Z's: $d_2 = -\frac{1}{\sqrt{2}}\left(\frac{4}{\sqrt{2}}\right) + \frac{1}{\sqrt{6}}(0) + \frac{1}{\sqrt{12}}\left(\frac{3}{\sqrt{12}}\right) + \frac{1}{\sqrt{20}}\left(\frac{-13}{\sqrt{20}}\right) + \frac{1}{\sqrt{30}}\left(\frac{-18}{\sqrt{30}}\right) = -3$, and so on.]

Through Equation 1 and Equation 2 you can get to the sample standard deviation s in one of two ways. The first way, which is the definition of the sample standard deviation, is to use the X's to compute d_1, …, d_6 and then get s through SSX. The second way is to start with the X's, then use Equation 1 to get Z_2, …, Z_6, then use these to get SSX and s through Equation 2. Since with the 5 numbers Z_2, Z_3, Z_4, Z_5, and Z_6 you can get SSX and s (and even get d_1, …, d_6) these 5 numbers contain the totality of the information that there is in the data set for measuring the spread in the population. This shows that the sample standard deviation s is really made up of 5 numbers and therefore really has 5 degrees of freedom. In a sense the information in d_6 has been absorbed into Z_2 through Z_6. So in total the Z's have 6 degrees of freedom, one degree of freedom in $Z_1 = \bar{X}$ for measuring the center of the population, and 5 degrees of freedom in Z_2, Z_3, Z_4, Z_5, and Z_6 for measuring the spread in the population.

The previous analysis extends to n data points, not just 6 data points, and the sample standard deviation will have $n-1$ degrees of freedom. Again $Z_1 = \bar{X}$, but now for the rest of the Z's the general Z_i, for $i = 2$ to $i = n$, is given by:

$$Z_i = \frac{X_1 + \cdots + X_{i-1} - (i-1)X_i}{\sqrt{i(i-1)}}.$$

Exactly what the Z's are is not important, what is important is that it is the case always that: $SSX = Z_2^2 + Z_3^2 + \cdots + Z_n^2$, no matter what the X's are. All the information in the data about spread in the population is contained in the numbers: Z_2, Z_3, …, Z_n, showing that SSX has exactly $n-1$ degrees of freedom. The property that the Z's have that make this happen is that if $Z_p = p_1X_1 + p_2X_2 + \cdots + p_nX_n$ and $Z_q = q_1X_1 + q_2X_2 + \cdots + q_nX_n$ and if $p_1q_1 + p_2q_2 + \cdots + p_nq_n = 0$, as in Equation 1, then this will make Z_p and Z_q uncorrelated. If $\{p_1, \ldots, p_n\}$ is considered to be a vector then the purpose of $\sqrt{i(i-1)}$ in the Z's is to give these vectors all a length of 1. For that reason square roots are ubiquitous in this paper. In advanced mathematics it is shown that lots of such sets like $\{p_1, \ldots, p_n\}$ and $\{q_1, \ldots, q_n\}$ exist. These types of sets will be used over and over again in the rest of the paper.

If we assume that X_1, X_2, …, X_n are independent and have normal distributions, with mean μ and standard deviation σ, then statisticians have proved, as mentioned previously, that the correlation between Z_i and Z_j is zero for $i \neq j$. This makes the Z's very different than the previous d's, the original deviations. Under these assumptions statisticians have proved that Z_2, Z_3, …, Z_n are normal with mean 0 and standard deviation σ, in addition to being uncorrelated. By definition when you square and add uncorrelated normal mean 0 random variables, such as in $Z_2^2 + Z_3^2 + \cdots + Z_n^2$, the chi-square distribution comes into play with degrees of freedom equal to the number of squared variables, here $n-1$. A chi-square distribution means that a formula has been found for calculating probabilities in this sums of squares situation just as a formula for the normal curve has been found and is used to calculate probabilities when the data is normal. Therefore chi-square distributions become relevant when you are measuring spread. When statisticians talk about degrees of freedom they are really talking about the degrees of freedom in the situation generated chi-square distribution. The original SSX, adds up squared correlated deviations and there is no simple distribution that can be applied to that situation. That is how and why one must go to the uncorrelated Z's to figure out the behavior of the original SSX.

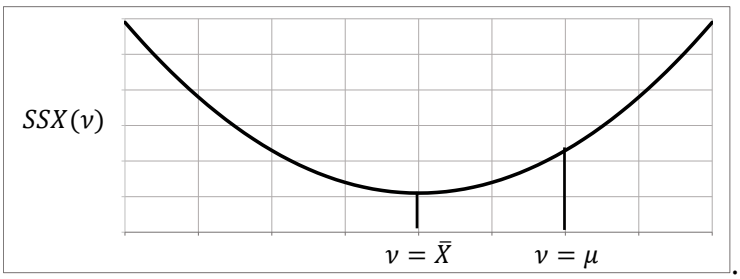
In the above, part of the data was used to measure the center of the population which is measured with $\bar{X}$. Suppose now that your Fairy God Mother told you the exact value of the population mean μ. Then it would not be necessary to use the data to estimate the center since you would know that the center is μ. That would signify that all of the data in the sample could then be used to measure the spread in the population. Let $Z_1 = X_1 - \mu$, $Z_2 = X_2 - \mu$, …, $Z_n = X_n - \mu$ and let:

$$\hat{\sigma} = \sqrt{\frac{SSX}{n}} = \sqrt{\frac{Z_1^2 + Z_2^2 + \cdots Z_n^2}{n}}.$$

This $\hat{\sigma}$ would measure the spread in the population and SSX would indeed have n degrees of freedom since it would use all of the n independent pieces of information in the X's for spread. The behavior of SSX would be determined by a chi-square distribution with n degrees of freedom if the X's are normal. Statisticians can prove that $\hat{\sigma}$ tends to correctly measure the size of the population standard deviation σ, meaning $Expected(\hat{\sigma}) \approx \sigma$ (see Bickel and Doksum, 1977, or Mood, Graybil, and Boes, 1974).

To see one last way why the quantity $n-1$ is used in the sample standard deviation, consider the following function: $SSX(v) = (X_1 - v)^2 + \cdots + (X_n - v)^2$, which is a function of the X's and a parameter v. The function is graphed in Figure 2.

Figure 2: The Function $SSX(v)$



The function $SSX(v)$, it can be proved, has a minimum at $v = \bar{X}$ so that $SSX(\bar{X}) < SSX(\mu)$ for the Fairy God Mother's μ. If $\hat{\sigma}$, which is based on $SSX(\mu)/n$, tends to have the correct size then an estimate of spread based on $SSX(\bar{X})/n$ would tend to be too small. This too small quantity can be made bigger by dividing by the smaller number $n-1$ rather than dividing by n, so that $SSX(\bar{X})/n < SSX(\bar{X})/(n-1)$. Statisticians have proved that $SSX(\bar{X})/(n-1)$ tends to have the correct size, and thus s, the sample standard deviation, which is based on this quantity satisfies $Expected(s) \approx \sigma$. The final thought for all of this is that taking $v = \bar{X}$ in $SSX(v)$ just fits the data, the X's, too well compared to $v = \mu$, making $SSX(\bar{X})/n$ too small. If your Fairy God Mother is mad at you and you do not have any knowledge of μ then you have no choice but to use s, but if your Fairy God Mother is a happy camper it will be more efficient to use $\hat{\sigma}$.

DEGREES OF FREEDOM IN REGRESSION

Suppose in a regression problem there are two predictor variables X_1 and X_2 and a response variable Y which is a function of X_1 and X_2 plus a random error term ε, i.e. specifically assume that $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. The values α, β_1, and β_2 are unknown parameters to be estimated from the data. A different error term ε turns up for each set of observations collected and it is assumed that the (set of) errors are independent. It is common in regression to assume, and we will assume, that the X-values are fixed and that the Y-values are the variables. The Y's that turn up will depend on the fixed X-values but also on the particular random error ε's that show up in the data. This is in contrast to the previous section where the X-values were the variables that could change, but are now fixed.

To estimate the parameters, a regression equation is fit to the data. Students are taught that estimates of the typical size of the prediction error when there are two predictor variables would have $n-3$ degrees of freedom, when there are n triplets of data (X_1, X_2, Y). These degrees of freedom for error usually appear for the first time in the Regression *ANOVA* Table. Why these are the degrees of freedom for the error term is the topic of this section. To understand what the degrees of freedom are, in this regression context, we again resort to transformations of the data as was done in the previous section. Now both the X's and the Y's will be transformed, but only Y will be considered a variable that can change.

For an example a cell phone company will want to predict the data requirements of its customers and will most likely use regression to make the predictions. Suppose the predictor variables are X_1 the family household income (in tens of thousands of dollars), and X_2 the size of the family. The dependent variable Y, which is a function of the X's (and the random error), will be the yearly number of gigabytes of cell phone data used by the household. Suppose a hypothetical data set of $n = 5$ triplets of data is $(X_1, X_2, Y) = (4,2,36), (6,2,41), (7,4,47), (11,6,83)$, and $(12,6,73)$. Below is the regression output for these triplets from Excel.

Figure 3: Regression Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.961558094
R Square	0.924593968
Adjusted R Square	0.849187935
Standard Error	8.062257748
Observations	5

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1594	797	12.26154	0.075406
Residual	2	130	65		
Total	4	1724			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12	10.81665383	1.1094	0.382787	-34.5403	58.5403
Income	3	4.163331999	0.7205	0.54601	-14.9134	20.9134
FamilaySize	5	7.059272862	0.7082	0.552189	-25.3736	35.3736

From the printout, the regression equation for this data set is: $Y = 12 + 3X_1 + 5X_2$. To see how well this equation fits this data we compute the root mean square prediction error ($RMSE$) in Table 2.

Table 2: Root Mean Square Error Calculations for the Regression Equation

		Actual Y	Predicted $Y = \hat{Y}$	Prediction Error	Squared Error
X_1	X_2	Y	$\hat{Y} = 12 + 3X_1 + 5X_2$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
4	2	$Y_1 = 36$	$\hat{Y}_1 = 34$	$e_1 = 2$	$e_1^2 = 4$
6	2	$Y_2 = 41$	$\hat{Y}_2 = 40$	$e_2 = 1$	$e_2^2 = 1$
7	4	$Y_3 = 47$	$\hat{Y}_3 = 53$	$e_3 = -6$	$e_3^2 = 36$
11	6	$Y_4 = 83$	$\hat{Y}_4 = 75$	$e_4 = 8$	$e_4^2 = 64$
12	6	$Y_5 = 73$	$\hat{Y}_5 = 78$	$e_5 = -5$	$e_5^2 = 25$
				Sum=0	Sum= $SSE = 130$

[SSE = Sum of Squares Error.], $RMSE = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{130}{5}} \approx 5.0990$.

A typical prediction error for the regression equation is approximately 5 which represents a typical number on the list of prediction errors: $Y - \hat{Y}$ (ignoring the minus signs). From Table 2 you can see: $e_1 + e_2 + e_3 + e_4 + e_5 = 0$, similar to the previous section. [This is true not just in this case but is true, in every regression data set, that the sum of the predictions errors is zero.] The prediction errors are therefore linked and in regression there is an even stronger relationship among the e's than there is for the d's in the standard deviation (where the last deviation was a function of the first $n - 1$ deviations). In regression when there are two predictors it always turns out that the last three prediction errors are always functions of the first $n - 3$ prediction errors. Since $n = 5$ for the cell phone data this tells us that for this data e_3, e_4 and e_5 are functions of e_1 and e_2. Given the X's in this data set it can be derived that always $e_3 = -2e_1 - 2e_2$, $e_4 = 2e_1 + 4e_2$, and $e_5 = -e_1 - 3e_2$, no matter what the Y's are. For the cell phone data this checks with Table 2 since: $e_3 = -2(2) - 2(1) = -6$, $e_4 = 2(2) + 4(1) = 8$, and $e_5 = -(2) - 3(1) = -5$. [If the data instead had been: $Y_1 = 36$, $Y_2 = 40$, $Y_3 = 54$, $Y_4 = 74$, and $Y_5 = 86$ with the same X's, then it is easy to check that the regression equation becomes: $Y = 10 + 4X_1 + 4X_2$ with $e_1 = 2$, $e_2 = -2$, $e_3 = 0$, $e_4 = -4$, and $e_5 = 4$. So again: $e_3 = -2(2) - 2(-2) = 0$, $e_4 = 2(2) + 4(-2) = -4$, and $e_5 = -(2) - 3(-2) = 4$.] These relationships state that there is no more additional information about prediction error in e_3, e_4 and e_5 than there already is in of e_1 and e_2.

The e's are obviously correlated since some are explicit functions of others, but for the degrees of freedom in SSE we can now rewrite SSE as

$$SSE = {e_1}^2 + {e_2}^2 + {e_3}^2 + {e_4}^2 + {e_5}^2 = {e_1}^2 + {e_2}^2 + (-2e_1 - 2e_2)^2 + (2e_1 + 4e_2)^2 + (-e_1 - 3e_2)^2.$$

This suggests that SSE has only two degrees of freedom despite the fact that originally it is computed with 5 numbers. Similar to what was done in the previous section this relationship can be made even more exact by making a slightly more complicated transformation of the e's which separates SSE into its uncorrelated parts. For these X's if you let $e_1' = (5/\sqrt{10})\, e_1 + 0e_2$ and $e_2' = \left(-15/\sqrt{30}\right)e_1 - \sqrt{30}e_2$ then it is always the case that:

$$SSE = {e_1}^2 + {e_2}^2 + {e_3}^2 + {e_4}^2 + {e_5}^2 = {e_1'}^2 + {e_2'}^2. \quad (3)$$

Here e_1' and e_2' have extracted and contain in totality all of the independent information that is available in the data for estimating the size of a typical prediction error, and in addition they are uncorrelated. All of the information available in e_1 to e_5 has been transferred to e_1' and e_2' so that SSE in this example is really a function of exactly two numbers demonstrating that SSE has exactly two degrees of freedom. For the original data $e_1' = \sqrt{10}$ and $e_2' = -2\sqrt{30}$ so that: $SSE = {e_1}^2 + {e_2}^2 + {e_3}^2 + {e_4}^2 + {e_5}^2 = (2)^2 + (1)^2 + (-6)^2 + (8)^2 + (-5)^2 = 130 = \left(\sqrt{10}\right)^2 + \left(-2\sqrt{30}\right)^2 = 10 + 120 = {e_1'}^2 + {e_2'}^2$.

The exact numbers involved in this example are not important. The important fact is that in every regression data set, a relationship such as in Equation 3 always exists. [This is true no matter what the Y's are, for if $Y_1 = 36$,

$Y_2 = 40$, $Y_3 = 54$, $Y_4 = 74$, and $Y_5 = 86$, with the errors for these Y's reported above, and which give $SSE = 40$, have $e_1' = \sqrt{10}$ and $e_1' = \sqrt{30}$ so that ${e_1'}^2 + {e_2'}^2 = 10 + 30 = 40$.]

In terms of understanding degrees of freedom in regression the formulas for the values of a, b_1, b_2, e_1', e_2' should be considered as transformations of the Y's. For these X's it can be calculated that:

$$a = Y_1 + .2Y_2 + .6Y_3 - .2Y_4 - .6Y_5,$$
$$b_1 = -.2Y_1 + (1/3)Y_2 - (4/15)Y_3 - (1/15)Y_4 + .2Y_5,$$
$$b_2 = .2Y_1 - (2/3)Y_2 + (13/30)Y_3 + (7/30)Y_4 - .2Y_5,$$
$$e_1' = \left(2/\sqrt{10}\right)Y_1 - \left(1/\sqrt{10}\right)Y_2 - \left(2/\sqrt{10}\right)Y_3 + 0Y_4 + \left(1/\sqrt{10}\right)Y_5,$$
$$e_2' = 0Y_1 - \left(1/\sqrt{30}\right)Y_2 + \left(2/\sqrt{30}\right)Y_3 - \left(4/\sqrt{30}\right)Y_4 + \left(3/\sqrt{10}\right)Y_5.$$

For example for the original cell phone data this checks as: $a = 36 + .2(41) + .6(47) - .2(83) - .6(73) = 12$, and so on. [It also checks for the other set of Y's just above.]

This transformation can be reversed and you can show that:

$$Y_1 = a + 4b_1 + 2b_2 + \left(2/\sqrt{10}\right)e_1' + 0e_2' = \hat{Y}_1 + e_1,$$
$$Y_2 = a + 6b_1 + 2b_2 - \left(1/\sqrt{10}\right)e_1' - \left(1/\sqrt{30}\right)e_2' \quad = \hat{Y}_2 + e_2,$$
$$Y_3 = a + 7b_1 + 4b_2 - \left(2/\sqrt{10}\right)e_1' + \left(2/\sqrt{30}\right)e_2' = \hat{Y}_3 + e_3,$$
$$Y_4 = a + 11b_1 + 6b_2 + 0e_1' - \left(4/\sqrt{30}\right)e_2' = \hat{Y}_4 + e_4.$$
$$Y_5 = a + 12b_1 + 6b_2 + \left(1/\sqrt{10}\right)e_1' + \left(3/\sqrt{30}\right)e_2' = \hat{Y}_4 + e_5.$$

The coefficients in these ten equations depend on the fixed X values, and such relationships exist in every regression data set. Since you can go from Y_1, Y_2, Y_3, Y_4, Y_5 to a, b_1, b_2 , e_1', e_2', and back again, the two sets (of five numbers) must have the same amount of information, just in a different form. The five degrees of freedom in Y_1, Y_2, Y_3, Y_4, Y_5 become one degree of freedom for the intercept a, two degrees of freedom total for the slopes, one each in b_1 and b_2, and two degrees of freedom in e_1', and e_2' for the error such that $SSE = {e_1'}^2 + {e_2'}^2$, and showing that SSE has precisely two degrees of freedom.

In general in regression with k predictors with the model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon$, and n sets of observations on both the independent and dependent variables, the n degrees of freedom in Y_1, Y_2, Y_3, …, Y_n become one degree of freedom for a, k degrees of freedom in b_1, b_2, b_3, …, b_k, and $n - k - 1$ degrees of freedom for SSE from e_1', e_2', e_3', …, e_{n-k-1}', such that $SSE = {e_1'}^2 + {e_2'}^2 + \cdots + {e_{n-k-1}'}^2$.
In advanced statistics it can be shown that $RMSE$, which divides SSE by n, tends to underestimate the size of a typical prediction error, in a similar way as to what happens in the sample standard deviation (the a, b_1, b_2, b_3, …, b_k, estimated from the data, fit the data too well to appropriately measure average error size). If we had our Fairy God Mother's help again and she told us the real values of the intercept α and the slopes β_1, β_2, β_3,…, then we would not have to use the data to estimate them and could devote all of the data to estimate the size of a typical error. In the absence of a Fairy God Mother, a more accurate estimate of the size of a typical prediction error, than $RMSE$, is given by the standard error of estimate s:

$$RMSE = \sqrt{\frac{{e_1}^2 + {e_2}^2 + \cdots + {e_n}^2}{n}}$$
$$s = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{{e_1}^2 + {e_2}^2 + \cdots + {e_n}^2}{n-k-1}} \sqrt{\frac{{e_1'}^2 + {e_2'}^2 + \cdots + {e'_{n-2}}^2}{n-k-1}}.$$

The RMSE from Table 2 was 5.0990, while the standard error of estimate s from the printout was 8.0623, and 8.0623 would be a more accurate measure of typical error size. The same symbol s is used both for the sample standard deviation and the standard error of estimate, the context tells you what situation you are in.

If the error terms ε in the regression data are independent and have normal distributions with mean zero then e_1', e_2', e_3', …, and e_{n-k-1}' are uncorrelated and have normal distributions with mean zero. In SSE you are squaring and

adding uncorrelated normal mean zero random variables and so the chi-square distribution with $n-k-1$ degrees of freedom makes an appearance, as is detailed in textbooks. When statisticians talk about degrees of freedom in this regression context also, they are referring to the degrees of freedom in the generated chi-square distribution.

THE DEGREES OF FREEDOM IN THE REGRESSION ANALYSIS OF VARIANCE TABLE

In a regression printout, such as in the Excel printout in Figure 3, one will often see an Analysis of Variance (*ANOVA*) Table of the form:

ANOVA	*SS*	*df*
Regression	$SSR=\sum\left(\hat{Y}-\bar{Y}\right)^2=b_1'^2+b_2'^2+b_3'^2+\cdots+b_k'^2$	k
Error (or Residual)	$SSE=\sum\left(Y-\hat{Y}\right)^2=e_1'^2+e_2'^2+\cdots+e_{n-k-1}'^2$	$n-(k+1)=n-k-1$
Total	$SSY(also\ called\ SST)=\sum(Y-\bar{Y})^2$	$n-1.$

That SSY has $n-1$ degrees of freedom is the topic of the first section. The b_i' have not been defined yet in the paper, so to do this note that SSR is based on $\hat{Y}$ and $\bar{Y}$. Since by definition $\hat{Y}=a+b_1X_1+b_2X_2+\cdots+b_kX_k$ for each set of X_1,…, X_k and since statisticians can prove that $\bar{Y}=a+b_1\bar{X}_1+b_2\bar{X}_2+\cdots+b_k\bar{X}_k$, it is the case that $SSR=\sum\left(\hat{Y}-\bar{Y}\right)^2=\sum[b_1(X_1-\bar{X}_1)+b_2(X_2-\bar{X}_2)+\cdots+b_k(X_k-\bar{X}_2)]^2$, so that SSR is a somewhat complicated function of b_1, b_2, b_3, …, b_k. In the cell phone data you can compute that $SSR=46b_1^2+52b_1b_2+16b_2^2$. For that data $b_1=3$ and $b_2=5$, which results in $SSR=1594$, as is verified in the Excel printout in Figure 3. In this form it is not exactly clear how SSR behaves or more specifically what distribution SSR has. As in previous situations this behavior is complicated since b_1 and b_2 are correlated in general. However here if we define $b_1'=2\sqrt{10}b_1+\sqrt{10}b_2$ and $b_2'=\sqrt{6}b_1+\sqrt{6}b_2$ then $SSR=b_1'^2+b_2'^2$ and it is also the case that b_1' and b_2' are uncorrelated and therefore separate SSR into its uncorrelated parts. This is not just good luck, such transformations exist in every regression data set. In the cell phone example then: $b_1'=11\sqrt{10}$ and $b_2'=8\sqrt{6}$ so that $SSR=b_1'^2+b_2'^2=\left(11\sqrt{10}\right)^2+\left(8\sqrt{6}\right)^2=1210+384=1594$. [This will be true no matter what the Y's are. If $Y_1=36$, $Y_2=40$, $Y_3=54$, $Y_4=74$, $Y_5=86$, then $b_1=4$, $b_2=4$, $SSR=46(4^2)+52(4)(4)+16(4^2)=1824$, $b_1'=12\sqrt{10}$, $b_2'=8\sqrt{6}$, so that $SSR=b_1'^2+b_2'^2=\left(12\sqrt{10}\right)^2+\left(8\sqrt{6}\right)^2=1440+384=1824$.] The b''s can also be expressed in terms of the Y's and for the X's in the cell phone data it can be shown that:

$$b_1'=-\left(2/\sqrt{10}\right)Y_1+0Y_2-\left(1/\sqrt{10}\right)Y_3+\left(1/\sqrt{10}\right)Y_4+\left(2/\sqrt{10}\right)Y_5$$
$$b_2'=0Y_1-\left(2/\sqrt{6}\right)Y_2+\left(1/\sqrt{6}\right)Y_3+\left(1/\sqrt{6}\right)Y_4+0Y_5.$$

It is possible to get b_1 and b_2 back from b_1' and b_2', so that the set $\{b_1, b_2\}$ and the set $\{b_1', b_2'\}$ contain the same information, just from a transformed perspective.

In the general case with k predictors there always exist b_1', b_2', b_3', …, b_k' which are functions of b_1, b_2, …, b_k, and such that $SSR=b_1'^2+b_2'^2+b_3'^2+\cdots+b_k'^2$. If the Y's have normal distributions then under the null hypothesis $\beta_1=0$, $\beta_2=0$, …, $\beta_k=0$, you can show that b_1', b_2', b_3', …, b_k' are all uncorrelated independent normal and mean zero, and when squaring them in SSR, SSR's behavior is obtained from a chi-square distribution with k degrees of freedom. The entirety of the information in the data about the stated null hypothesis is contained in b_1', b_2', b_3', …, b_k'. This hypothesis test involves both SSR and SSE, the signal in SSR compared to the noise in SSE. The F-statistic in the printout in Figure 3 is $F(Stat)=[SSR/k]/[SSE/(n-k-1)]$ which is an appropriately scaled ratio of chi-square distributions and has an F-distribution. An F-distribution simply means that someone found a formula for calculating the probability that $F(Stat)$ will be a small number, a big number, or something in between. In an F-test these probabilities can be used to decide if $F(Stat)$ provides evidence for or against that null hypothesis.

In *ANOVA* the quantity SSR, in total, has k degrees of freedom, one for b_1, one for b_2, …, and one for b_k. The Y_1, Y_2, …, Y_n have n degrees of freedom. The sum of squares SSR and SSE combined have $k+(n-k-1)=n-1$ degrees of freedom. The remaining one degree of freedom is for the intercept a. The intercept is not represented in this type of *ANOVA* Table but is represented in other forms of Regression *ANOVA* Tables (such as in Draper and

Smith, 1966). Finally in the above table the word "Residual" is another name for "Error". Residual represents what is left of Y after you make the prediction, i.e., after you take away the predicted value $\hat{Y}$ from Y you get a residual, in other words: $Residual = Y - \hat{Y}$.

The theory of regression writes regression models in terms matrices and the well-developed mathematics of matrices was used to compute the funny numbers in the paper. This theory can ultimately be used to prove the properties specified in this paper and put forth in statistics books.

DEGREES OF FREEDOM IN ANALYSIS OF VARIANCE PROBLEMS

Analysis of Variance Problems consist of One Way ANOVA, Two Way ANOVA without replications, Two Way ANOVA with replications with or without interaction terms, Three Way ANOVA with or without two way interactions with or without three way interactions, Four Way ANOVA, etc. The same logic as in the previous sections of the paper apply to this case as well. Consider a Two Way ANOVA problem without replications with the data and model presented in Table 3. A Two Way ANOVA has two different factors that affect the data and each factor will have multiple levels. For the hypothetical data in Table 3, on weights of two week old piglets, the two factors are gender and type of feed.

Table 3: Weights of Piglets

Factor A:		Factor B: Gender	
Feed	Weight	Female	Male
Type	Feed Type 1	$Y_{11} = \mu + \alpha_1 + \beta_1 + \varepsilon_{11} = 13$	$Y_{12} = \mu + \alpha_1 + \beta_2 + \varepsilon_{12} = 19$
	Feed Type 2	$Y_{21} = \mu + \alpha_2 + \beta_1 + \varepsilon_{21} = 9$	$Y_{22} = \mu + \alpha_2 + \beta_2 + \varepsilon_{22} = 11$
	Feed Type 3	$Y_{31} = \mu + \alpha_3 + \beta_1 + \varepsilon_{31} = 14$	$Y_{32} = \mu + \alpha_3 + \beta_2 + \varepsilon_{32} = 12$

For a Two Way ANOVA without replications, with *r* levels of Factor A and *c* levels of Factor B, the general form of the *ANOVA* Table is as follows:

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*
Rows (Factor A)	***SSA***	***r*-1**	***SSA*/(*r*-1)**	***MSA/MSE***
Columns (Factor B)	***SSB***	***c*-1**	***SSB*/(*c*-1)**	***MSB/MSE***
Error	***SSE***	**(*r*-1)(*c*-1)**	***SSE*/[(*r*-1)(*c*-1)]**	
Total	***SSY***	***rc*-1**		

The data in Table 3 has $r = 3$ and $c = 2$. Below is *ANOVA* Printout from Excel for that data:

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows (A: FeedType)	***SSA* = 36**	**2**	**18**	**2.25**	**0.307692**	**19**
Columns (B: Gender)	***SSB* = 6**	**1**	**6**	**0.75**	**0.477767**	**18.5128**
Error	***SSE* = 16**	**2**	**8**			
Total	***SSY* = 58**	**5**				

If we let $\bar{\bar{Y}} = \frac{Y_{11}+\cdots+Y_{32}}{6}$, $\bar{Y}_{i\cdot} = \frac{Y_{i1}+Y_{i2}}{2}$, and $\bar{Y}_{\cdot j} = \frac{Y_{1j}+Y_{2j}+Y_{3j}}{3}$, then $SSY = \sum\sum\left(Y_{ij} - \bar{\bar{Y}}\right)^2$, $SSA = \sum\sum\left(\bar{Y}_{i\cdot} - \bar{\bar{Y}}\right)^2$, $SSB = \sum\sum\left(\bar{Y}_{\cdot j} - \bar{\bar{Y}}\right)^2$, and $SSE = \sum\sum\left(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{\bar{Y}}\right)^2$.

Consider the transformation of this data defined by:

$$\bar{\bar{Y}} = Y_{11}/6 + Y_{21}/6 + Y_{31}/6 + Y_{12}/6 + Y_{22}/6 + Y_{32}/6 = 13,$$
$$a_1' = 2Y_{11}/\sqrt{12} - Y_{21}/\sqrt{12} - Y_{31}/\sqrt{12} + 2Y_{12}/\sqrt{12} - Y_{22}/\sqrt{12} - Y_{32}/\sqrt{12} = 18/\sqrt{12},$$

$$a_2' = 0Y_{11} + Y_{21}/2 - Y_{31}/2 + 0Y_{12} + Y_{22}/2 - Y_{32}/2 = -3,$$
$$b_1' = Y_{11}/\sqrt{6} + Y_{21}/\sqrt{6} + Y_{31}/\sqrt{6} - Y_{12}/\sqrt{6} - Y_{22}/\sqrt{6} - Y_{32}/\sqrt{6} = -\sqrt{6},$$
$$e_1' = 2Y_{11}/\sqrt{12} - Y_{21}/\sqrt{12} - Y_{31}/\sqrt{12} + 2Y_{12}/\sqrt{12} - Y_{22}/\sqrt{12} - Y_{32}/\sqrt{12} = -\sqrt{12},$$
$$e_2' = 0Y_{11} + Y_{21}/2 - Y_{31}/2 + 0Y_{12} - Y_{22}/2 + Y_{32}/2 = -2.$$

Given $\bar{\bar{Y}}$, a_1', a_2', b_1', e_1', and e_2' you can get back Y_{11}, Y_{21}, Y_{31}, Y_{12}, Y_{22}, and Y_{32}. This transformation has the property that no matter what the Y's are, in this three levels for A and two levels for B example, it is the case that: $SSA = {a_1'}^2 + {a_2'}^2$, so that SSA has two degrees of freedom, $SSB = {b_1'}^2$, so that SSB has one degree of freedom, and $SSE = {e_1'}^2 + {e_2'}^2$, so that SSE has two degrees of freedom. For the above Y's this is verified since: $SSA = \left(18/\sqrt{12}\right)^2 + (-3)^2 = 36$, $SSB = \left(-\sqrt{6}\right)^2 = 6$, and $SSE = \left(-\sqrt{12}\right)^2 + (-2)^2 = 16$. If the Y's are normal, the behavior of SSA, SSB, and SSE can be determined from chi-square distributions with the reported degrees of freedom.

The above ANOVA Table has four rows of effects. A complete Three Way ANOVA will have nine rows of effects (Scheffé 1959, 123) including main effects and interactions. Statisticians have proved that such transformations as above exist for every ANOVA problem. In ANOVA, if SSZ has m degrees of freedom there will always be m transformations of the Y's, z_1', …, z_m', such that $SSZ = {z_1'}^2 + \cdots {z_m'}^2$. Using this method it can be proved that the degrees of freedom for the main effects, the two way interactions, the three way interactions, the measurement errors and so on, are those degrees of freedom that are detailed in textbooks. Systematic procedures exist for finding the appropriate transformations in each ANOVA case (Haberman, 1974, pp. 150-153) and they then can then be used to determine the behavior of SSA, SSB, etc. through the chi-square distribution. While we have only done proof by example in this paper, all of this has been made mathematically precise in the general case.

CONCLUSION

The words "Degrees of Freedom" are used all the time in statistics. For most people who have had a statistics class, the idea of degrees of freedom still remains a mystery. This paper shows where degrees of freedom come from at a relatively elementary level, for both the standard deviation and in regression and finally in ANOVA. The paper does this by considering the sample average, the sample standard deviation, and regression coefficient estimates as transformations of the data. By appropriately defining transformations of the data, the number of degrees of freedom in each situation is revealed and from that their properties are discerned. If the data comes from a normal distribution then squaring things, as in the standard deviation, leads to chi-square distributions. When statisticians talk about degrees of freedom they are really referring to the degrees of freedom of the chi-square distribution that each statistical situation brings about. Everything you ever wanted to know about degrees of freedom was discussed in the paper.

REFERENCES

Bickel, P., and Doksum, K. (1977), *Mathematical Statistics*, San Francisco: Holden-Day.
Draper, N., and Smith, H. (1966), *Applied Regression Analysis*, New York: John Wiley and Sons
Haberman, S. (1974), *The Analysis of Frequency Data*, Chicago: University of Chicago Press
Mood, A., Graybil, F., and Boes, D. (1974), *Introduction to the Theory of Statistics* (3rd ed), New York: McGraw-Hill.
Rao, C.R. (1973), *Linear Statistical Inference* (2nd ed), New York: John Wiley and Sons.
Scheffé, Henry (1959), *The Analysis of Variance*, New York: John Wiley and Sons.

Kenneth Sutrick, Ph.D., teaches statistics in the Bauernfeind College of Business at Murray State University, Murray, Kentucky. He has a B.A. in Mathematics from the University of Wisconsin-Madison and a Ph.D in Statistics from the University of California-Berkeley. He spends a lot of time thinking about how to teach the difficult topics of introductory statistics. His research interests include portfolio theory, and options.