THE IMPORTANCE OF STATISTICAL MODELING IN DATA ANALYSIS AND INFERENCE

Derrick K. Rollins, Sr.
Iowa State University • Ames, Iowa 50011

Unquestionably, the most widely used construct in the science and engineering literature in regards to data analysis and statistical inference is "error bars." In an extensive search in the top 25% of physiology journals between Jan. 1 and March 31, 2014 (n = 703), Weissgerber, et al.[1] found at least one "bar graph" in about 86% of the articles. I did a study a few years ago in one particular journal that produces a large number of experimental articles and examined every article over a 20-year period and found that 75% had error bars in a plot or table. There are basically two types of "error bars" used in this literature. The first one, that shall be called "Form 1," is simply the sample mean, $\bar{x}$, plus or minus ($\pm$) the sample standard deviation, s, or $\bar{x} \pm s$. The second one, that shall be called "Form 2," is similar and given by $\bar{x} \pm s\left(\sqrt{n}\right)^{-1}$, where n is the sample size. Note that $s\left(\sqrt{n}\right)^{-1}$ is the estimated standard error of the mean. Error bars appear to be a construct outside of the statistical community since they are not even mentioned in the most widely (perhaps any) used statistical textbooks.

To my knowledge, there are basically two kinds of discussions in the science and engineering literature. One discussion centers on differentiating the two forms and their application.[2,3] The other one is about a better way of presenting information from data when the sample size is small.[1] However, neither of these discussions strongly suggests the elimination of the use of error bars completely on the grounds that they are not sound on a statistical modeling basis. Thus, the objective of this article is to show that the use of error bars should be eliminated completely in statistical data analysis and inference on the basis of soundness. This article will support this contention using a statistical modeling approach and showing that error bars either do not comply with the model or have very low statistical significance when they do comply, which essentially makes them useless. More specifically, Form 1 is not an interval estimator for any population parameter and while Form 2 can be considered an interval estimator for a population mean, its level is too small to be useful. This contention will be defended on the basis of sound statistical analysis and inference from the perspective of the statistical model. Moreover, I contend that when sound statistical modeling is practiced and data analysis and inference are consistent with the model, error bars will be clearly seen as erroneous and unfounded. This contention is supported in this article for several common cases: one population, two populations, one-way ANOVA (say p populations), multifactor ANOVA (say pq populations for two-way ANOVA), and regression. By taking a clearly specified modeling approach, this work will

Derrick K. Rollins, Sr. *is a professor in Chemical & Biological Engineering and Statistics. He received a B.S. degree in chemical engineering from the University of Kansas, and M.S. in statistics, as well as M.S. and Ph.D. in chemical engineering, from The Ohio State University. He previously worked as a process engineer for the E.I. Du Pont Chemical Company and as a faculty intern for the 3M Company. His research areas include glucose monitoring, modeling, and control to help people with diabetes control blood sugar better, and (medical-, bio-, and material-) informatics.*

unquestionably show that error bars have no sound basis for use in statistical analysis and inference and will provide statistically sound alternative approaches in all these applications that can be taught to students in introductory statistics courses, and applied to experiments in undergraduate education and all types of research including academia and industry. Moreover, this article is written at the undergraduate introductory level of an engineering statistics course and assumes familiarity and experience with the cases and methodologies covered by the scope of this article.

ONE POPULATION CASE

The experiment

Before taking on the task of writing a statistical model, the population(s) should be clearly defined, the goal(s) of the study clearly stated, and the experiment and data collection clearly articulated. The One Population Case presented here has two population parameters for some attribute of interest, the mean (μ) and the standard deviation (σ) with N items in the population. These quantities are commonly unknown. The experiment is to randomly select n items from the population and obtain the value of the attribute for each one.

Statistical model

Defining x_i as the value of this attribute for the ith selection (i = 1, . . ., n) the statistical model is given as follows:

$$x_i = \mu + \varepsilon_i \qquad (1)$$

where ε_i is independently distributed $\left(\overset{\text{indep}}{\sim}\right)$ with a mean of 0 and variance of σ^2 for i = 1, . . . , n and written as

$$\varepsilon_i \overset{\text{indep}}{\sim} \left(0, \sigma^2\right) \quad \forall i \qquad (2)$$

The statistical model gives a detailed description of how the experimenter believes the data will behave. For soundness I proposed the following practices. First, that a detailed statistical model be given when statistical inference is implied or assumed in data collection studies. Second, that articles submitted for review containing data studies give detailed statistical models as a requirement to be accepted for publication. In my experience I have found that often the experimenters cannot clearly specify the statistical model. This inability is an indication that the person does not completely understand what he/she is doing and this leads to statistically unsound practices and analyses.

A pictorial description of this case is given in Figure 1. As illustrated, the sampling from the population is a "Random Sample (RS)" of size n. An RS of size n is a sample of n items taken from a population that has the same probability of being selected as any other sample of size n. The statistical model [Eqs. (1) and (2)] indicates that sampling is independently and identically distributed, *i.e.*, an RS.

The goal of this case is to evaluate hypotheses for one or both of the population parameters (μ and σ). For example, one might be interested to know if $\mu > \mu_0$ or if $\sigma < \sigma_0$ where μ_0 and σ_0 are fixed values.[4]

Inference for σ

From Eqs. (1) and (2), the variance of x_i $\left(\sigma^2_{x_i}\right)$ is $\sigma^2 \forall i$ and the variance of $\bar{x}$ $\left(\sigma^2_{\bar{x}}\right)$ is σ^2/n. Thus, σ is a measure of spread of the samples about their true mean of μ. This understanding of σ is critical. Equally critical is the understanding that its value is the result of all sources that cause variability about the mean. These sources include the differences of the numerical value of the items in the population as well as any sources of experimental (*i.e.*, sampling) error in obtaining the value of any sample. Thus, the size of σ is a reflection of the quality of the data since the likelihood of larger experimental errors mean a larger value of σ.

It is critical to understand the difference between a sample quantity and a population quantity. This is often confused in the engineering and scientific literature and this confusion leads to misunderstanding and erroneous conclusions. More specifically, s is not σ. While it is a point estimate of σ it can be a highly inaccurate one. The likelihood that it is a highly inaccurate estimate is the situation that is most often found in engineering and science literature, that is, when n is small.

For this one population case, Forms 1 and 2 error bars are $\bar{x} \pm s$ and $\bar{x} \pm s\sqrt{n}$, respectively. As shown, Form 1 gives an estimated spread (s) about an estimated mean ($\bar{x}$). Form 2, which will be discussed in more depth later, is a poor interval estimate of the true mean, μ, and does not convey information of the spread of samples about their mean. Thus, it is not considered further in this discussion about spread. When n is small a scatter plot of all the data is more informative than Form 1 as it reveals all the information.[1] When n is sufficiently large, formal statistical inference (*i.e.*, confidence intervals and hypothesis testing) for σ will be far superior

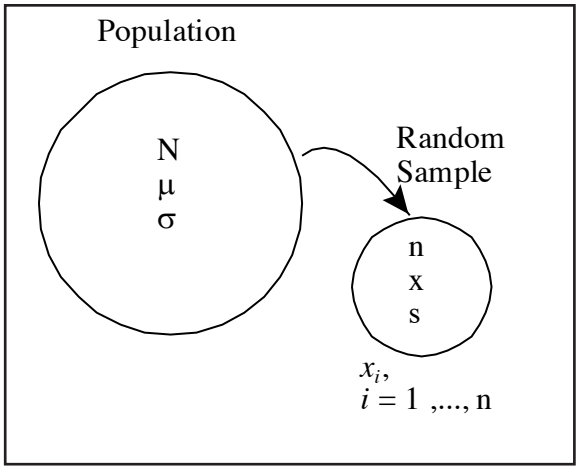

Figure 1. *Representation of the One Population Case.*

than presenting $\bar{x} \pm s$ for inference of the spread about μ. Under normality, for this one population case, a 100(1- α)% confidence interval for σ is

$$s\sqrt{\frac{(n-1)}{X^2_{\alpha/2,n-1}}},\quad s\sqrt{\frac{(n-1)}{X^2_{1-\alpha/2,n-1}}} \tag{3}$$

where $X^2_{\alpha,n-1}$ is the 100(1- α)th percentile of the X^2 distribution with n –1 degrees of freedom. Note that formal statistical inference can always be done regardless of the sample size but when the sample size is small, for example, confidence intervals will not likely be very precise (*i.e.*, are likely to be very wide) and the true level of spread is likely to be better assessed in the scatter plot of the data.[1]

2.4 Inference for μ

The point estimator for μ is $\bar{x}$. The standard error of $\bar{x}$ is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and its point estimator is $s_{\bar{x}} = s/\sqrt{n}$. As any estimator, $\bar{x}$ is unreliable for small n. For statistical inference regarding μ in this one population case, the use of error bars cannot provide a statistically sound analysis that is consistent with the model given by Eqs. (1) and (2). Essentially, error bars are interval estimates. However, Form 1, $\bar{x} \pm s$, is not an interval estimate for any parameter in the population. Actually, it is not an interval estimate for any type of parameter. An interval estimate for some parameter θ will contain the estimate of θ, $\hat{\theta}$, and the standard error of $\hat{\theta}$, $\sigma_{\hat{\theta}}$ or the estimated standard error of $\hat{\theta}$, $s_{\hat{\theta}}$. Form 1, $\bar{x} \pm s$, does not meet this requirement since s is not the estimated standard error of $\bar{x}$. Furthermore, while Form 2, $\bar{x} \pm s/\sqrt{n}$, does meet this requirement as an interval estimate for μ, its level of reliability is so small that any sound analysis would not even consider it. The type of interval estimator that has measures of both precision and reliability is a confidence interval. Under normality for the above model, a 100(1- α)% confidence interval for μ is[4]

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{2}{\sqrt{n}} \tag{4}$$

where $t_{\alpha,n-1}$ is the 100(1-α)th percentile of the student-t distribution with n -1 degrees of freedom. For Form 2, $t_{\alpha/2,n-1} = 1$. For n = 5 and 10, Form 2 gives 63% and 66% confidence intervals, respectively, levels that are too low to be of value.

In summary for this one population case, neither Form 1 or Form 2 have merits in assessing spread about the mean —this should be assessed using scatter plots with all the data when n is small and using formal statistical analyses when n is large enough for the estimator s to be reliable. Form 1 has no value in statistical inference for μ while Form 2 is an interval estimate for μ that is too low to be of any value and not worth considering.

TWO POPULATION CASE

The experiment

The goal of this case is to evaluate hypotheses for the Population 1 parameters (μ_1 and σ_1) in relation to the Population 2 parameters (μ_2 and σ_2).[4] For example, one might be interested to know if $\mu_1 > \mu_2$ or is $\sigma_1 < \sigma_2$. Defining x_i as the i^{th} random selection from Population 1 (i = 1, . . . , m) and y_j as the j^{th} random selection from Population 2 (j = 1, . . . , n), *i.e.*, the experiment is to take a random sample of size m from Population 1 and a random sample of size n from Population 2.

The statistical model

The statistical model is:

$$x_i = \mu_1 + \varepsilon_i \tag{5}$$

$$y_j = \mu_2 + \varepsilon_j \tag{6}$$

where

$$\varepsilon_i \overset{indep}{\sim} (0,\sigma_1^2) \quad \forall i \tag{7}$$

$$\varepsilon_j \overset{indep}{\sim} (0,\sigma_2^2) \quad \forall i \tag{8}$$

$$\varepsilon_i \text{ indep } \varepsilon_j \quad \forall i,j \tag{9}$$

A pictorial description of this case is given in Figure 2.

Inference for σ

Forms 1 and 2 error bars for populations 1 and 2 are, respectively: $\bar{x} \pm s_1$, $\bar{x} \pm s_1/\sqrt{m}$; $\bar{y} \pm s_2$, $\bar{y} \pm s_2/\sqrt{n}$. Form 1 error bars should not be used to assess a difference in spread between the two populations for the same reasons mentioned above in the one population case. For small sample sizes, scatter plots that use all the data are most informative.[1] For sufficiently large sample sizes for both populations, formal statistical inferential procedures should be used. For example, under normality, for this two-population

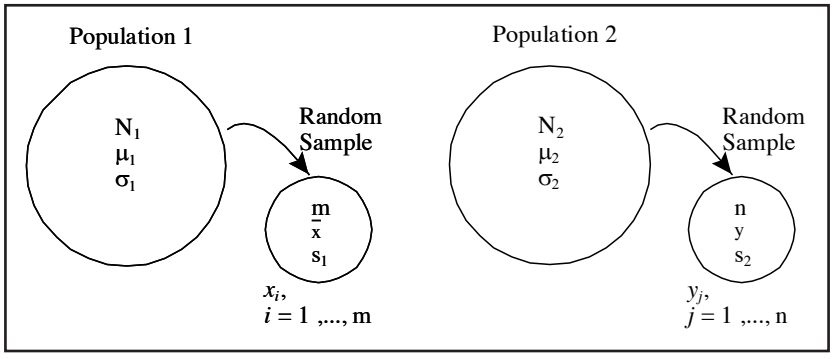

Figure 2. *Representation of the Two Population Case.*

case, a 100(1-α)% confidence interval for σ_1 / σ_2 is[4]

$$\frac{s_1}{s_2}\sqrt{\frac{1}{f_{\alpha/2,m-1,n-1}}},\ \frac{s_1}{s_2}\sqrt{f_{\alpha/2,n-1,m-1}} \qquad (10)$$

where $f_{\alpha,m-1,n-1}$ is the 100(1-α)th percentile of the f distribution with m – 1 numerator degrees of freedom and n – 1 denominator degrees of freedom. Note that formal statistical inference can always be done regardless of the sample sizes but when the sample sizes are small, for example, this confidence interval will not likely be very precise and the true difference in spread is likely to be better assessed by comparing the scatter plots of the data.

Inference for μ

As in the one population case, Form 2 error bars are not statistically sound interval estimates for the two population case for the same reason. When the sample sizes are small, scatter plots using all the data are better informative tools for assessing differences in population means.[1] In all other cases, formal statistical inference will be best. Under normality for the two population model, a 100(1-α)% confidence interval for $\mu_1 - \mu_2$ is[4]

$$\bar{x} - \bar{y} \pm t_{\alpha/2,v}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} \qquad (11)$$

where

$$v = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{\left(\frac{s_1^2}{m}\right)^2}{m-1} + \frac{\left(\frac{s_2^2}{n}\right)^2}{n-1}} \qquad (12)$$

ONE-WAY ANOVA CASE

The experiment

The one-way ANOVA case is an extension of the two population case to I populations with the same variance σ^2. In ANOVA the term "Factor" is used as a common description for the I populations. For example, a study may consist of determining the best of four formulations for an explosive mixture. The factor could be described as Formulation with I = 4 levels or populations. For this case J samples will be taken from each population and is called "the number of replicates." Thus, the total number of runs, n_t, is IJ. The n_t are completely randomized, *i.e.*, the IJ samples are taken in random order. The measurement on each sample is called the "response."

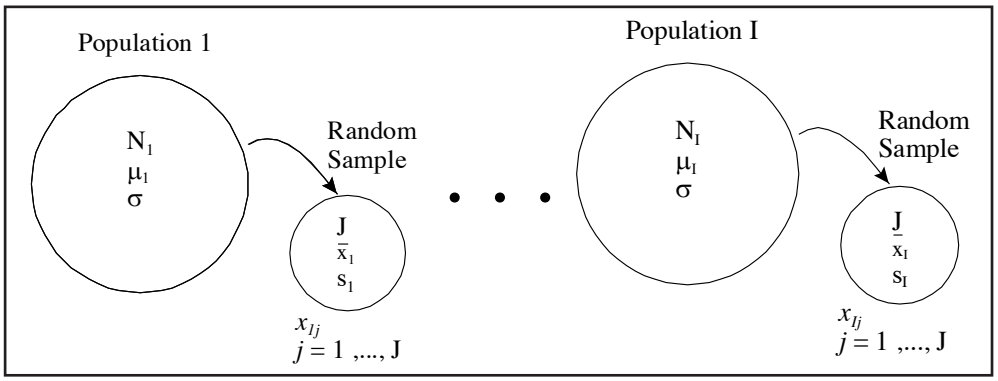

Figure 3. *Representation of the One-Way ANOVA Case.*

The statistical model

As previously, the statistical model is a mathematical description of the behavior of the response under the conditions of the experiment and stated assumptions. In this case the statistical model is written as follows[4]

$$x_{ij} = \mu_i + \varepsilon_{ij} \qquad (13)$$

where

$$\varepsilon_{ij} \overset{indep}{\sim} N(0,\sigma^2) \quad \forall ij \qquad (14)$$

x_{ij} is the measured response for the ith level of the factor on the jth replicate, μ_i is the true mean of the response for the ith population or level of the factor, ε_{ij} is the random deviation of the response from μ_i of the jth replicate and has a normal distribution with constant variance σ^2 for all n_t samples. A pictorial description of this model is depicted in Figure 3.

Inference

When the model is based on an assumption of constant variance any analysis or procedure that seeks to convey otherwise is in direct violation of this assumption. Thus, to use error bars to convey a variation in spread for different populations makes no sense in this context. However, the constant variance assumption should be assessed and leading statistical packages, such as Minitab, have tests for assessing this assumption.

Thus, in one-way ANOVA, inference is strictly focused on the means. In this case the null hypothesis is H_0: all the μ_i's are equal versus the alternative hypothesis, H_a: at least two of the μ_i's are not equal. Small sample size is typically not an issue in One-Way ANOVA because the constant variance assumption allows "pooling" of the information to determine an estimate for σ. Thus, an informal analysis in One-Way ANOVA is likely unnecessary.

When H_0 is rejected at a specified level of significance,

another procedure is needed to assess differences between the means. The use of Form 2 error bars is not sound for the reason mentioned previously, a low level of reliability, but also because multiple tests for pairs of means reduces the overall level of reliability even further which is often not reflected or appears to be misunderstood by those who use error bars. The sound statistical approach to maintaining high reliability in multiple inference for pairs of means is called "multiple comparisons" and uses a paired confidence interval approach that focuses on maintaining a high overall or simultaneous level of confidence. A widely accepted procedure for this model is the T-method (T is for Tukey).[4] In this method, a set of confidence intervals for the difference between two population means for all pairs of means are obtained (*e.g.*, $\mu_1 - \mu_2$). Taken together, the overall level of confidence that all of them are true is 100(1-α)%. More specifically, a 100(1-α)% simultaneous confidence interval for pairs of $\mu_i - \mu_j$, $i \neq j$, for all possible pairs, is

$$\bar{x}_i - \bar{x}_i \pm Q_{\alpha, I, I(J-1)} \sqrt{\frac{MSE}{J}} \qquad (15)$$

where

$$MSE = \frac{s_1^2 + \ldots + s_I^2}{1} \qquad (16)$$

and $Q_{\alpha,I,I(J-1)}$ is the 100(1-σ)th percentile of the studentized range distribution with I numerator degrees of freedom and I(J-1) denominator degrees of freedom. All confidence intervals that do not contain 0 conclude the means to be significantly different. This is the statistically sound way to do this analysis. No approach using Forms 1 and 2 error bars can be statistically sound.

Example

A set of data taken under the One-Way ANOVA case is shown in Table 1. The factor in this study is formulation and has four levels (I = 4) and three replicates (J = 3) for each level of the factor. The response is yield. A bar chart with Form 2 error bars for this data set is shown in Figure 4. This chart is a typical way that error bars are shown in the literature. More specifically: (1) they rarely identify the type of error bar that is on the chart; (2) the type of parameter of focus (means or spread) is not usually stated; (3) the level of significances is not ever given and; (4) the conclusions are not stated.

Notwithstanding, the sample standard deviations are based on a sample size of 3 and are thus, highly uncertain. However, it is common practice for error bars to be constructed from a sample size around 3.

Figure 5 is a plot that the statistical software package Minitab produced for this data set. Its superiority for statistical inference is immediately obvious over Figure 4. More specifically: (1) the type of parameter is stated, *i.e.*, μ ; (2) the significance level is stated, *i.e.*, 95%; (3) the type of inference is given, *i.e.*, confidence intervals; (4) the type of confidence intervals, *i.e.*, Tukey's and simultaneous, *i.e.*, based on Eqs. (15) and (16) and; (5) the conclusions are stated for each pair of means. If error bars were presented as in Figure 5, they would be exposed and their impotence in formal statistical inference would be revealed. Thus, its survival in the literature, to a large extent, has been essentially the lack of this exposure.

TABLE 1
Blend Data in the One-Way ANOVA Case

	Formula			
	1	2	3	4
Yield	25.6	25.2	20.8	31.6
	24.3	28.6	26.7	29.8
	27.9	24.7	22.2	34.3

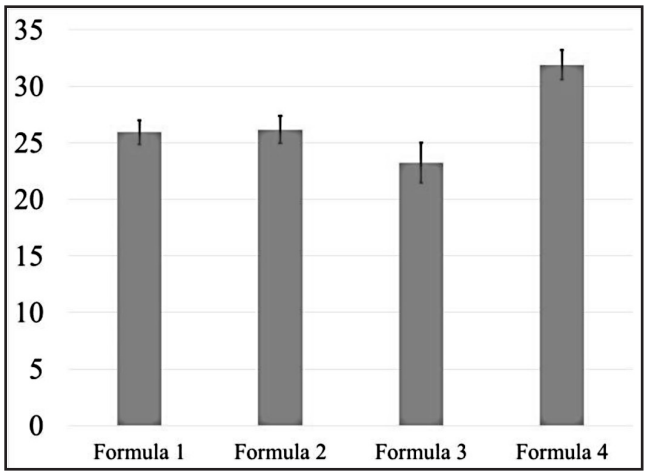

Figure 4. *Average yield for the four formulas (the data from Table 1) in the One-Way ANOVA case. This bar graph is a typical way that error bars are displayed in the literature. The type of error bar is not identified. The type of parameter of interests (μ or σ) is not stated. The level of significances is not given and conclusions are not given.*

MULTIFACTOR ANOVA CASE

The experiment

Multifactor ANOVA is an extension of one-way ANOVA to more than one factor. The number of populations is the product of the levels for the factors. The one presented here is a Two-Way ANOVA case as this allows the populations to be visualized in two dimensions (see Figure 6). The number of levels for factors A and B are I and J, respectively. Thus, in this case there are IJ populations or "treatment combinations." In this case the number of replicates for each treatment combination is K and is constant. Here we consider the case where K > 1. Thus, n_t is IJK. The n_t are completely randomized, *i.e.*, the IJK samples are taken in random order. As in

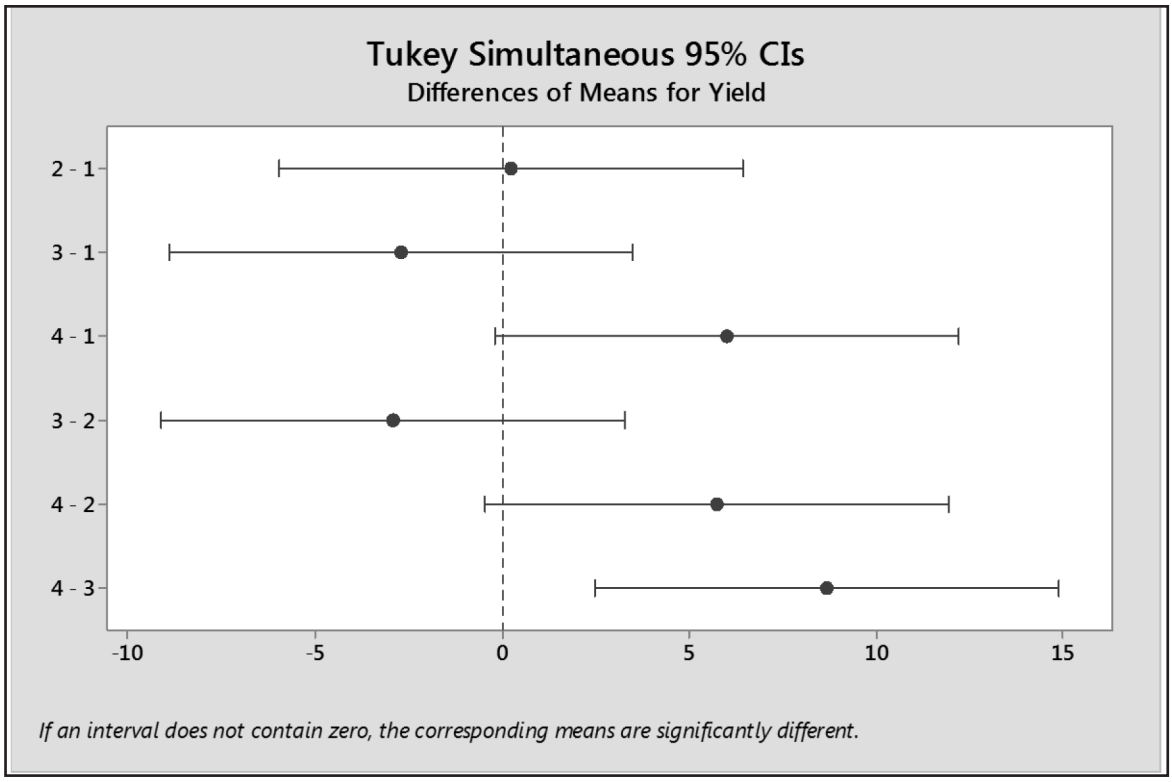

Figure 5. *Graphical results from Minitab for the formula data set (Table 1) in the One-Way ANOVA case. The information here is far superior to the information presented in typical error bar plots as shown by a comparison with Figure 4.*

one-way ANOVA, σ^2 is assumed to be the same for each of the IJ populations. The factors are fixed which means that they each have a set of specific populations that were not subject to random selection from a larger set of populations.

Two-Way example

Before giving the Two-Way model, it will be informative to give the example for this case since it puts the model into a real context. This experiment was actually run in the classroom of a statistics course that I taught recently. It is a broad jumping experiment. Factor A is person and it has three levels (I = 3), that is, three people. This factor is fixed. Factor B is activity, it is also fixed and it has two levels (J = 2). The first level of the activity was standing behind the line and jumping. The second level was coming up to the line, closing the eyes, patting the head, rubbing the stomach and turning around three times. After they stopped I would align them so that they were pointed straight ahead. They could actually be far in front or behind the line before jumping. The number of replications for each treatment combination (IJ = 6) is K = 3. Thus, n_t is IJK = 18. These 18 trials were run completely in a random order. The response is the distance a person lands from the marked line.

The statistical model

The statistical model for this experiment is given as follows:

$$x_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad (17)$$

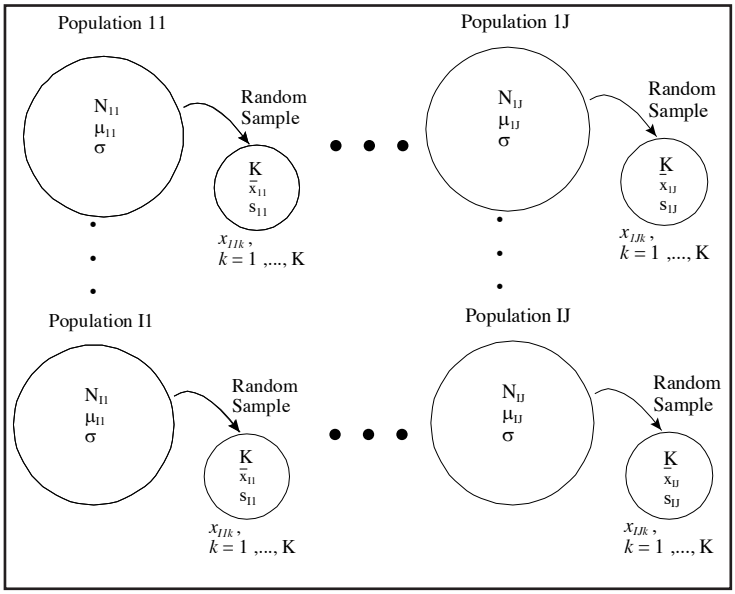

Figure 6. *Representation of the Two-Way ANOVA Case.*

where

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad (18)$$

$$\sum_{i=1}^{3} \alpha_i = 1, \sum_{j=1}^{2} \beta_j = 0, \sum_{i=1}^{I} \gamma_{ij} = 0, \sum_{j=1}^{J} \gamma_{ij} = 0 \qquad (19)$$

$$\varepsilon_{ijk} \overset{indep}{\sim} N(0, \sigma^2) \quad \forall ijk \qquad (20)$$

TABLE 2
Results of the Two-Way Experiment with the Run Order Given

		Person					
		1		2		3	
		Distance (in)	Order	Distance (in)	Order	Distance (in)	Order
Activity	1	67.8	11	87.3	17	74.8	15
		70.0	10	80.3	1	76.0	12
		68.5	4	87.0	5	69.5	7
	2	54.4	14	76.5	6	91.4	16
		53.8	2	87.1	8	84.0	13
		57.1	18	56.3	3	64.9	9

TABLE 3
ANOVA Table from Minitab for the Two-Way Experiment

Analysis of Variance for Distance					
Source	DF	SS	MS	F	P
Person	2	1040.04	520.02	6.71	0.011
Activity	1	72.28	172.28	2.22	0.162
Person*Activity	2	375.22	187.61	2.42	0.131
Error	12	930.36	77.53		
Total	17	2517.91			

x_{ijk} is the distance landed from the line for the ith person, on the jth activity, for the kth replicate; μ_{ij} is the true mean distance from the line for the ith person on the jth activity; μ is the grand mean distance from the line; α_i is the main effect for person i; β_j is the main effect for activity j and; γ_{ij} is the interaction effect for ith person on the jth activity. Thus, the response can be describe mathematically as the grand mean (μ), plus an adjustment for person (α_i), plus an adjustment for activity (β_j), plus a correction for the interaction between person and activity (γ_{ij}) plus random deviation ($\varepsilon_{i\,jk}$) attributed to phenomenon like measurement error. As a scientist or engineer seeking to effectively apply statistical inference to the results of their study, he/she should be able to describe the response mathematically in this way and to write out the statistical model in the details given here. If the person cannot do this, it is likely that they will have difficulty understanding the results statistically and in soundly applying statistical inference. On the other hand, when the person can do this it is not likely that will apply unsound methodologies like the use of error bars. The ability helps to keep them on a solid foundation for sound statistical inference when all applications, methodologies and analyses are scrutinized by this understanding.

Hypotheses

As in the case in one-way ANOVA all the hypotheses are focused on the means for multi-factor ANOVA since the variance is assumed to be constant for all the populations. Hypotheses are tested in the order from the highest order interaction to the lowest order interaction, then to the main effects lastly, depending on the testing results at each level. The hypotheses for this study and the proper order of testing are[4]:

H_{0AB}:$\gamma_{ij} = 0\forall$ij versus H_{aAB}: at least one $\gamma_{ij} \neq 0$; H_{0A}: α_i=0$\forall$i versus H_{aA}: at least one $\alpha_i \neq 0$ and; H_{0B}: $\beta_j = 0\forall$j versus H_{aB}: at least one $\beta_j \neq 0$.

Results

The responses obtained for this study are given in Table 2. An ANOVA table (a table that is commonly used to provide the results for the effects in a study,[4] *e.g.*, in this case the two main effects and the interaction effect) from Minitab is given by Table 3. The significance level is set at 0.05. The significance level (α) is the type 1 error rate, that is, the probability of rejecting a null hypothesis (H_0) when it is true .[4] Note that Table 3 gives the P-value for evaluating the significance of each effect. The P-value for each case is determined from the results (*i.e.*, the sampled data). The results sufficiently support statistical significance when the P-value is less than α. Starting with the interaction hypothesis test (HT), since its P-value = 0.131 is greater than 0.05, the evidence is not strong enough to conclude that the interaction is significant (*i.e.*, H_{0AB}:$\gamma_{ij} = 0\forall$ij is not rejected in favor of H_{aAB}: at least one $\gamma_{ij} \neq 0$). Conversely for the factor Person, its P-value = 0.011 is less than 0.05 so this factor is found to be statistically significant (*i.e.*, H_{0A}: α_i=0$\forall$i is rejected in favor of H_{aA}: at least one $\alpha_i \neq 0$). Finally, the factor Activity, with a P-value of 0.162, is not found to be statistically significant (*i.e.*, H_{0B}: $\beta_j = 0\forall$j is not rejected in favor of H_{aB}: at least one $\beta_j \neq 0$).

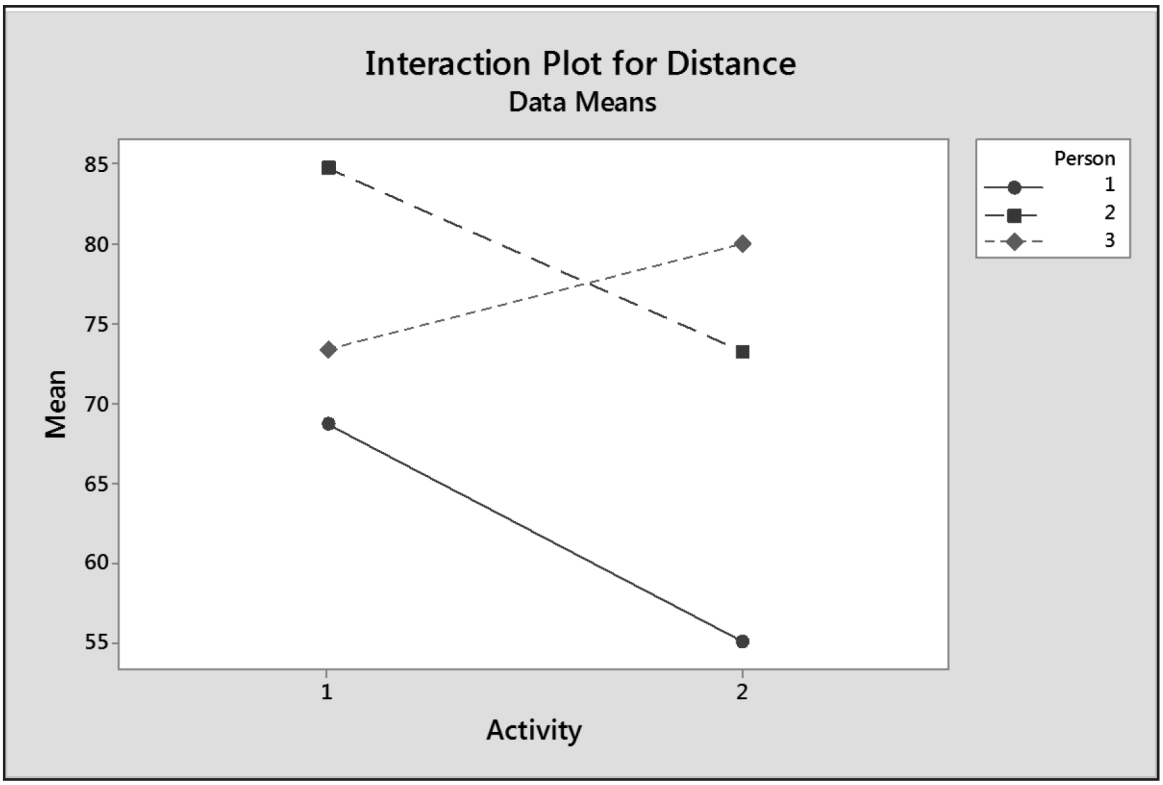

Figure 7. *Interaction plot for the Two-Way example from Minitab.*

The estimate for σ, s, is equal to $\sqrt{\text{MSE}} = 8.8$ inches which is quite high. A large value of s adversely affects the ability to detect significant effects, *i.e.*, to have high statistical power. This appears to be the case here as an informal analysis will reveal. An interaction plot from Minitab is given in Figure 7. This plot shows clearly that the interaction is significant. Thus, both factors are concluded to be significant. This analysis illustrates the importance of informal analysis when power is low.

An analysis checking the assumption of equal variances will reveal that it does not appear to be true for this experiment. However, since ANOVA is very robust for departures from this assumption when the factors are fixed, as in this case, this is not a concern. Minitab has a formal test for checking the assumption of equal variances and provides a discussion of its strengths and weaknesses under certain conditions as well as references on procedures that can be used when this assumptions is not adequate. This is true for many statistical software tools.

Typically, in this type of study, error bars would be put on bar graphs for each level of the factor and bar graphs for the interactions would not be shown. Since the type of parameter being assessed is the true mean, Form 2 error bars would be the only applicable ones. In addition to the limitation of assessing interaction, the weaknesses of error bars described in the one-way case would also be revealed here and in any ANOVA study.

REGRESSION

Statistical model

While the discussion in this article will be given in the context of linear regression (LR) it is applicable to regression in general as it relates to error bars. Regression is a statistical methodology to obtain a fitted model for the response (y) on a set of explanatory variables (x) when change in y is affected by the change in x in a continuous manner even though the actual changes in x may be discrete. A general model for LR with k carriers is as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i = \mu_{y_i/x_i} + \varepsilon_i \qquad (21)$$

where

$$\varepsilon_i \overset{\text{indep}}{\sim} N\left(0, \sigma^2\right) \quad \forall i \qquad (22)$$

$x_i = [x_{1i} \ldots x_{ki}]^T$, β_0 is an unknown non-carrier parameter, β_l (l = 1, …, k) is an unknown carrier parameter associated with carrier x_l, and the common assumptions follow for ε_i as given by Eq. (22). Note that the constant variance assumption is also applicable to the regression model. The objective is to obtain values for the unknown parameters that agree with the model assumptions and give the "best" fit based on the least squares criterion given below:

$$\text{Minimize SSE} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \qquad (23)$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}$ and $\hat{\beta}_0$ and $\hat{\beta}_I$ (l = 1, …, k) are the estimators of β_0 and β_I (l = 1, …, k), respectively, that satisfy Eq. (23) and are called the least squares estimators.

Regression example

The example for regression comes from a case of fitting a model to real data presented to me by a colleague. The approach used by him is one that appears often in the literature in modeling data using regression. The fit of the model by my colleague is shown in the plot on the left in Figure 8. The bars in this plot are not error bars but 95% confidence intervals for mean values of y at the values of x shown. These confidence intervals were obtained using replicated data at the given values of x. While these are confidence intervals, they could have been error bars, as they too have appeared in this literature in a similar manner. While I commend my colleague for attempting to apply a legitimate statistical methodology, it was misapplied in violation of the model and critical features of the model were not exploited to obtain the best fit and sound inference.

The confidence intervals for $\mu_{y|x_i}$ were obtained using replicated values of the response at each x_i (*i.e.*, time) as shown in the plot. Since the variation of these values is different at each x_i, the width of the confidence intervals vary as shown. Form 2 error bars would have varied similarly but their width would have been narrower because their level of confidence is much lower than 95% as mentioned above. Nonetheless, determining confidence intervals in this manner, or forming error bars, do not follow the regression model, more specifically, its assumption of constant variance which would give much smaller estimated standard errors for $\hat{\mu}_{y|x_i}$ and thus, tighter 95% confidence intervals.

To obtain inference that complies with the regression model, I refit the data using linear regression techniques. The results for this fit are also shown in Figure 8 in the plot on the right that was generated by Minitab. First, since the responses do not vary linearly with time, I fit a third order polynomial. The model assumptions for this fit held fairly well except that the residuals $(e_i = y_i - \hat{y}_i)$ were somewhat serially correlated which is not surprising since the data were sequentially collected. (For addressing serial correlation in regression to improve parameter estimation see Reference 5. Similarly, for exploiting serial correlation to improve prediction, see the pre-whitening approaches in References 6 and 7.) However, observing that the serial correlation was not too strong I continued with the analysis and obtained 95% confidence bands for $\mu_{y|x}$ and 95% prediction bands for future predictions of y given $x\left(y_{future|x}\right)$. Prediction bands are wider than confidence bands due to the additional uncertainty in the variability about the estimate of the mean, *i.e.*, $\hat{\mu}_{y|x}$. For a given x, the upper and lower values of the bounds give the limits of the confidence intervals or the prediction intervals at that value of x. These limits are narrower on the average than those in the left plot in Figure 8 due to using the pooled estimate of σ. General equations for a 100(1-α)% confidence interval for $\mu_{y|x}$ and a 100(1-α)% prediction interval for $y_{future|x}$ are given by Eqs. (24) and (25), respectively[4]:

$$\hat{\mu}_{y/x} \pm t_{\alpha/2,n-(k+1)} S_{\hat{\mu}_{y/x}} \tag{24}$$

$$\hat{\mu}_{y/x} \pm t_{\alpha/2,n-(k+1)} \left[S^2_{\hat{\mu}_{y/x}} + s^2 \right]^{1/2} \tag{25}$$

where $s_{\hat{\mu}_{y|x}}$ is the estimated standard error of $\hat{\mu}_{y|x}$. The importance of complying with the regression model is clearly illustrated in Figure 8 for sound statistical analysis and inference which error bars cannot provide. Thus, the practice of using error bars in regression should not even be considered on the basis of noncompliance to the model that will lead to erroneous statistical analysis and inference.

CONCLUDING REMARKS

The use of error bars is extensive in the science and engi-

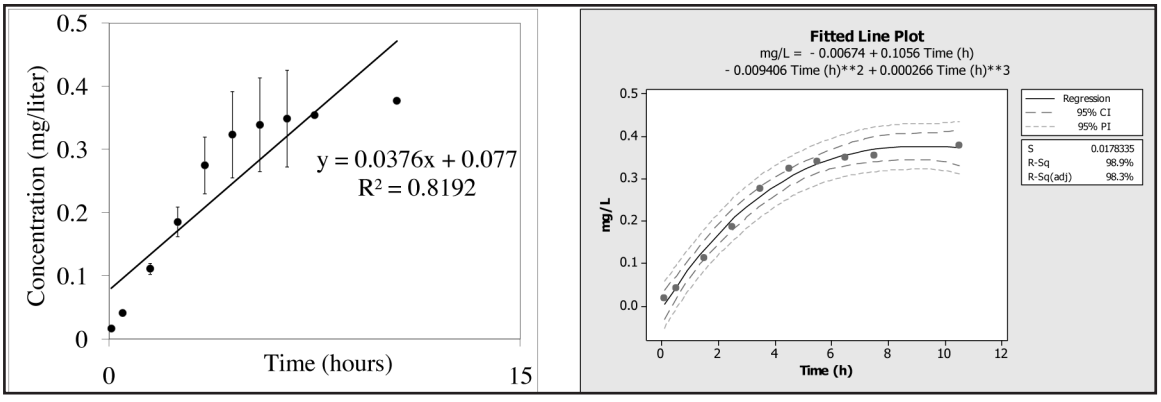

Figure 8. *The regression example with the original results on the left and the Minitab results on the right. For the plot on the left, 95% confidence intervals for* $\mu_{y|x}$ *are determined using replicated data at one value of* x_i (i.e., *time). The plot on the right uses a pooled estimate of* σ *in compliance with the regression model to obtain 95% confidence bands for* $\mu_{y|x}$ *and 95% prediction bands for* $y_{future|x}$.

neering literature. However, as shown in this work using a modeling based approach, error bars do not provide sound statistical data analysis and inference. For situations of low power, commonly due to small sample size but can also occur from large spread about the mean (*i.e.*, a large σ), informal statistical analysis such as scatter plots are more informative than error bars.[1] While Form 1 error bars reflect some information for spread about the mean, formal statistical procedure for inference about σ, such as confidence intervals or hypothesis testing, are more informative and thus, superior. Similarly, while Form 2 error bars are interval estimators for means, their levels are too low to be of any value, which drops even more for the overall level in multiple comparison situations. Thus, the construction of confidence intervals for means should not be done with error bars and should follow sound statistical practice of complying with the model assumptions, giving the level of confidence or significance, identifying the population parameter and simultaneously holding at a specified level in the context of multiple comparisons.

The importance of modeling in sound statistical inference was illustrated in several cases. These cases are common ones in data collection and analysis and showed directly how the use of error bars did not comply with these models as well as their incorrect use or impotence in sound statistical inference. The reality of error bars is that there is no sound justification for their use in statistical analysis and inference and their use should stop. This work proposes requiring a clearly and accurately stated statistical model whenever data are collected and statistical analysis and inference are done. This practice, when applied soundly, will result in the elimination of error bars, as illustrated by the cases in this work. Thus, as more researchers use statistical modeling in data collection studies, the quality of data analysis and statistical inference will improve and the use of error bars will vanish, as it should.

ACKNOWLEDGMENTS

I want to thank Andrea Lowe for her assistance in researching the use of error bars in scientific journals and Yong Mei for his help in preparing the manuscript for submission.

REFERENCES

1. Weissgerber, T.L., N.M. Milic, S.J. Winham, and V.D. Garovic, "Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm," *PLoS Biol*, **13**(4): e1002128. doi:10.1371/journal. pbio.1002128 (2015)
2. Davies, H.T., Describing and Estimating: Use and Abuse of Standard Deviations and Standard Errors," *Hospital Medicine*, **59**, 327 (1998)
3. Curran-Everett, D, and D.J. Benos, "Guidelines for Reporting Statistics In Journals Published by the American Physiological Society," *Physiological Genomics*, **18**, 249 (2004)
4. Devore, J.L., *Probability and Statistics For Engineering and the Sciences,* 7th ed., Duxbury Press (2007)
5. Neter, J., W. Wasserman, and M.H. Kutner, *Applied Linear Regression Models*, Homewood, Ill: Richard D. Irwin, Inc. (1983)
6. Box, G.P., and G.M. Jenkins, *Time Series Analysis: Forecasting and Control,* Revised ed. Oakland: Holden-day (1976)
7. Rollins, D.K., N. Bhandari, S. Chin, T.M. Junge, and K.M. Roosa, "Optimal Deterministic Transfer Function Modeling in the Presence of Serially Correlated Noise," *Chemical Engineering Research and Design*, **84**(A1): 9-21 (2006) ❐

ChE errata

Due to a production error, there is one symbol missing in the paper "Level Control..." by Larry K. Jang, published in the Fall 2016 issue of *CEE*. The letter "f" is missing in the final print. Eq. (1) on Page 245 should appear like

$$q = f(x,h) = C_v(x)\sqrt{\frac{\Delta P_{valve}}{S.G.}} = kx\sqrt{h} \qquad (1)$$