WHAT DID THAT PROFESSOR SAY?
Statistics made easy

We are surrounded by numbers every day. You may not realize it, but statistics plays a large role in our daily lives as well. Weather forecasting takes numbers and makes predictions about the weather based on weather models. Disease models for predicting turfgrass diseases do a similar service. Based on numbers related to temperature, humidity and leaf wetness, these models can forecast the startup of a turfgrass disease. We know that pest control products are tested for their effectiveness to control pests. Statistics are behind every medical study and batting average you hear about. Soon we will be bombarded with those political voter polls.

Statistics are sets of mathematical equations that are used to analyze what is happening in the world around us. It is a science of decision making. It is a science of "chance" or "probability." It is the science of collecting, organizing, and interpreting data whether it is numerical or non-numerical. We live in an information and technological age where we have everything at our finger tips. H.G. Wells, the father of science fiction, predicted that statistical thinking would be as necessary for daily living as reading and writing. Statistics may seem intimidating at first, but it is not once you develop a clear understanding of this simple subject.

BASIC UNDERSTANDING OF TERMS

Before we start, a discussion and understanding of some basic terms are needed. *Descriptive statistics*[are used to describe sets of numbers such as plants heights achieved due to applications of fertilizers. Researchers can organize these numbers into tables and graphs called *requency distributions* (the frequency a number may occur due to a factor involved). The following data set illustrates measurements of plant heights in centimeters after a fertilizer application). We will use this data to help us define some terms.

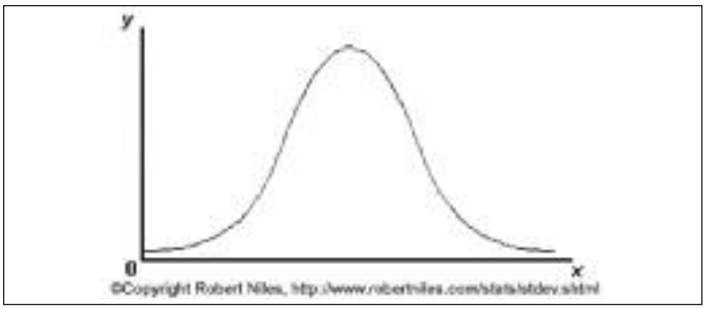
Plant Heights (cm) due to Fertilizer Applications				
10	14	11	12	15
15	12	13	14	13
12	8	12	9	10
13	11	12	8	10
9	16	7	11	9

As we look that this simple data set, we can determine a **median**, a **mean**, and **a standard deviation**. The median is the measurement that lies in the middle of the data, at the 50th percentile. In this example, it is 12 (range is 7-16). At times, it is better to express the median rather than the average (also known as the mean, see below), especially if the data contains outliers. The median could be a better indicator of true center especially when NBA salaries are being discussed.

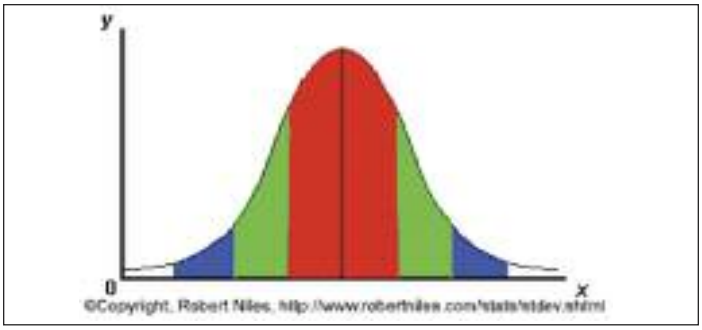
Plant Height (cm)	Frequency	Percent	Percentile
7	1	4	4
8	2	8	12
9	3	12	24
10	3	12	36
11	3	12	48
12	5	20	68
13	3	12	80
14	2	8	88
15	2	8	96
16	1	4	100
Totals	**25**	**100**	

The mean is simply the average (plant height x frequency observations = 286 cm / 25 frequency observations = 11.44 cm) for the data set. The standard deviation (SD = 2.38) indicates the average difference individual data varies from the mean; how concentrated the data are around the mean. So why is this important? Without standard deviation, you cannot get a feel for how close the data are to the mean or whether the data are spread out over a wide range. Without standard deviation, you cannot compare two data sets effectively. Two data sets can have the same mean, but vary greatly in the concentration of data around the mean; therefore different standard deviations.

The **distribution** of a data set can be a graph of all values and their frequency of occurrence. One of the most common distributions is called the normal distribution or *bell-shaped curve*displaying numerical data in a symmetrical curve.



The center of the bell is the mean and most of the data is usually centered on the mean.



The red area represents this data and one standard deviation +/- from the mean, 68% of the data (34% on either side of the average). The green area represents two standard deviations +/- from the mean or 95% of the data (red plus green) under the curve. The blue area then represents three deviations +/- from the mean or 99% of the data. Since every set of data has a different mean and standard deviation, an infinite number of normal distribution curves exist.

Confidence intervals (CI), usually set by the researcher, establish a level of confidence or reliability to an end result based on some treatment perhaps to a human being or plant in repeatable trials. The CI is represented by a percentage, so when we say, "we are 95% confident that the result of this herbicide application will provide 98% control of dandelion," we express that 95% of the observations will hold true. In practice, confidence intervals are typically stated at the 95% confidence level. However, they can be shown at several confidence levels like, 68%, 95%, and 99%. When a research trial is conducted, the confidence level is the complement of the respective level of significance, i.e. a 95% confidence interval reflects a significance level of 0.05, referred to as alpha (α). The level of confidence is often dependent on the number of observations with more observations yielding a higher level of confidence.

When data is collected, researchers typically look for something unusual or out of the ordinary and often ask if this is significantly different from a norm. Will it or does this happen with a very small probability of happening just by chance? **Least Significant Difference** (LSD) is a measure of significance usually with a level of significance (α= 0.05) denoted as $LSD_{\alpha= 0.05}$=0.05 or $LSD_{0.05}$. We will revisit the use of this term when we show an example of a data table and bar graph.

EXPERIMENTAL DESIGNS

How an experiment is designed can make the difference between the collection of good data and bad data. The objective of experiments is to make comparisons of *treatments* that will support a thought or hypothesis about an area of interest. Treatments can include the applications of fertilizers or pesticides, the incorporation of a cultural practice or the evaluation of disease resistant turfgrass cultivars or combinations thereof. While comparisons of treatments are important, so are comparisons to an untreated control to determine the true effects of each treatment if nothing was being applied. The untreated control establishes a baseline for comparison. Collecting good data and then applying the proper data analysis is important for drawing or making appropriate conclusions about the experiment.

In experimental designs, data (measurements/observations) are usually subject to various, uncertain external factors. Treatments and full experiments are usually repeated, *replications*, to help identify any sources of variation, to better estimate the true effects of the treatments thereby strengthening the reliability and validity of the experiment. Statistically, replications help to reduce experimental error due to unknown or uncontrollable factors (i.e. variations in soils). Replicating treatments within an experiment is as important as repeating entire experiments to see if results can be repeated with confidence. **Randomization** is also an important component to experimental design. One way to minimize bias in an experiment is to randomize treatments. This will become clearer as we look at some experimental designs.

Two common experimental designs that you may hear of in a seminar or conference presentation are illustrated below.

Complete Randomized Block Designs are one of the simplest, most common experimental designs for field trials. Here, you may be looking at the effects of one type of treatment, i.e. herbicide effectiveness. Treatments can be replicated three, four or more times dependent on the type of trial it is. Disease trials tend to have more replications due to the high variability among treatments from replication to replication. Treatments also remain in single blocks.

Complete Randomized Block Design							
Replicate 1	7	4	6	1	3	5	2
Replicate 2	6	4	1	7	5	3	2
Replicate 3	5	7	2	3	1	4	6

You will note that seven treatments are completely randomized in each of three replications or blocks. The treatment numbers can correspond to a treatment list.

Treatment No.	Treatments
1	Untreated control
2	Herbicide A, Rate 1
3	Herbicide A, Rate 2
4	Herbicide B, Rate 1
5	Herbicide B, Rate 2
6	Herbicide C, Rate 1
7	Herbicide C, Rate 2

Again, it is the randomness of the treatments that will eliminate bias of plot location within each block along with replicating the treatments that will help to increase reliability of the data.

Split Plot Designs are a special experimental design when several factors are being evaluated or some constraint (i.e. turfgrass species) prevents you from using a complete randomized block design. A variable could be the application of fungicides to test disease control on these specific turfgrass species. The diagram above demonstrates a split plot design.

Split Plot Design

Replicate 1	A 5	A 2	A 1	A 4	A 3	B 1	B 3	B 5	B 4	B 2
Replicate 2	B 3	B 5	B 1	B 2	B 4	A 4	A 3	A 2	A 1	A 5
Replicate 3	A 4	A 3	A 5	A 1	A 2	B 2	B 1	B 3	B 5	B 4

In many cases you need to fit the experiment into existing resources, like an established stand of grass. You will note that blocks A and B (i.e. two turfgrass species) are planted in blocks as a constraint of the experimental design, but are randomized within each replication. Within each replication, fungicide treatments are then randomized within each species. Treatment 1 may correspond to an untreated control, while treatments 2 through 5 may correspond to four different fungicides.

Additional experimental designs are available dependent on the number of factors being looked at; however, the more factors (i.e. species, fertilizers, pesticides, cultural practices, etc.), the more difficult it is to analyze, make comparisons, and draw conclusions.

ANALYZING THE DATA

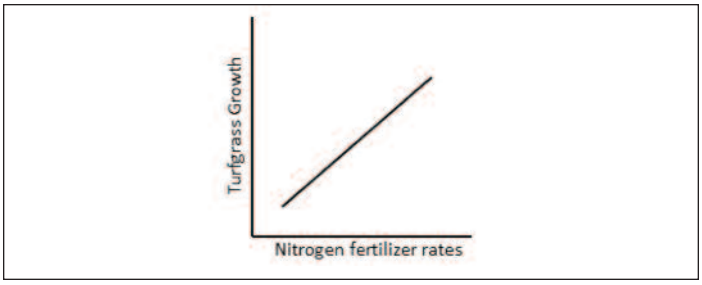
After all the data has been collected, the choice of analysis is just as important as the experimental design. This is often considered the black box of statistics. The wrong analysis can lead to wrong conclusions. Researchers need to ask themselves this, "Will I be able to legitimately and correctly answer the questions that I set out to answer after the data has been analyzed?" **Regression and Correlation** can be used to test a cause and effect relationship and how well that relationship is correlated. An **Analysis of Variance** (ANOVA) can be used to test the effectiveness of one product to another and how well that data may fit a regression line.

Regression is all about relationships answering questions like, "Does nitrogen fertilizer cause turfgrasses to grow taller?" Here we can relate two variables like fertility and growth and understand that we may observe a positive slope on a graph—turfgrasses will grow taller with increasing rates of nitrogen

Continued on page 44

Continued from page 10

fertilizer (X-axis equaling increasing rates of nitrogen fertilizer and the Y-axis equaling turfgrass growth).



However, we also need to ask, "Is nitrogen fertilizer the only factor that can increase growth?" The answer is obviously "no." External variable such as temperature and rainfall can influence results as well. So we can see that statistical relationships are not so clear cut and analyses try to find the best fit (the slope of the line) for this relationship.

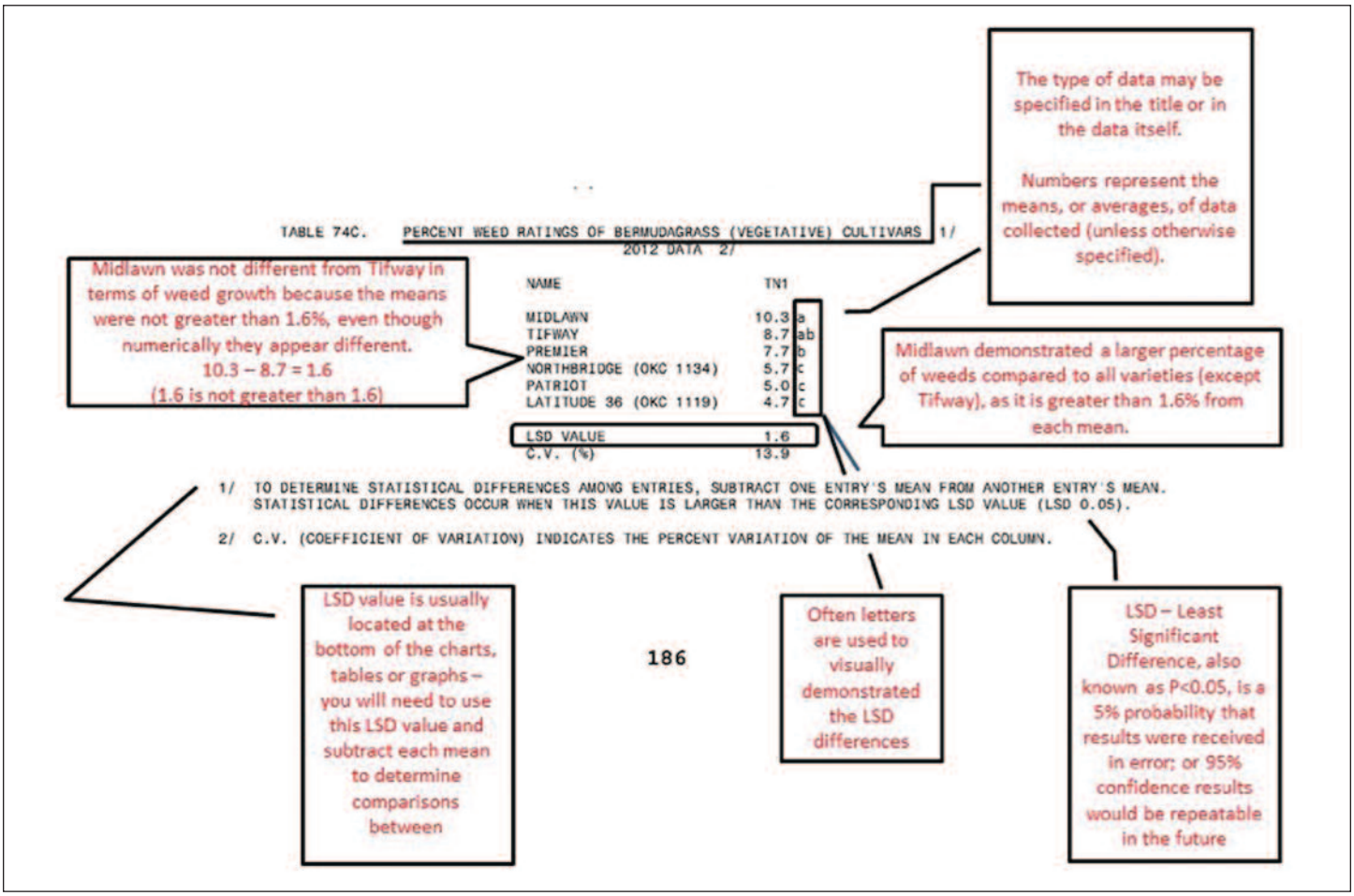
ANOVA is used to analyze differences or equality between treatment means. ANOVAs are useful for comparing two or more means for statistical significance. Significance between means is often determined by a threshold value such as the Least Significant Difference as one measure.

Analysis of data can be very confusing, drawn out and beyond the scope of this article. Those of us in Plant Sciences often consult with statisticians to aid in the analysis of large data sets. Let's leave this up to the experts.

EXAMPLES OF TABLES AND CHARTS & WHAT TO LOOK FOR

Understanding data tables becomes an easier task now that you understand some terms like the mean, standard deviation and least significant difference. The following example comes from the National Turfgrass Evaluation Program website. All tables should be titled; columns labeled and have some indication of significance between means.

This example shows a data table for weed ratings in some bermudagrass cultivars. The numbers listed under TN1 are means of three replications of percent weed ratings. Several text boxes explain much of the information on the data table; however, the most important question to ask, "Are there any differences, significant differences? " You will noted that the Least Significant Difference (LSD) value is 1.6 If the differences between means is greater than 1.6, then you will see a different lower case letter adjacent to that mean. It also specifies that the LSD is an LSD set at 0.05 or a 95% confidence level. Means with the



same lettering adjacent to it, are statistically equal (even if numerically appearing different).

Understanding Bar Graphs can appear to be easier than large data tables. They can present data in a cleaner, more simplified format; however, some cautions should be pointed out. First look at the vertical or y-axis and determine what measurement is being labeled and the scale. All scales should start at "0", but sometimes do not. Look at the units on the scale. Unit interval (unit interval of 1 versus a unit interval of 20) may tell you that differences in the bars are not as great as they may appear.

Just as data tables should, bar graphs should have some indication of mean separation and significance. Bars labeled with the same letters are equal to one another. Those with different letters (A versus B) are significantly different from each other. Bar graphs should be titled as well and have both axis labeled.

Data tables and bar graphs can be used to present supporting data for conclusions being made. Researchers will sometimes present large data tables and cluttered bar graphs that will cause you as a viewer to lose interest simply because you are unable to keep up with what is being said by trying to follow the numbers.

When a presenter displays data in a table or chart, there should be a reason to show such data other than just showing the numbers. When a table or chart is used in a PowerPoint, the presenter should explain all of the parameters of the information: what is it showing, define the numbers, explain the X and Y axis on a graph, point out and explain the level of significance and where significance exist. Highlighting areas of interest to make a point, or two at the most, is often best where large tables are used, but often not followed. This becomes difficult for the participant to pick up on the key points and often interest is lost in the presentation. Most often it is best to express large amounts of data as text statements rather than showing the numbers. For it's the results or conclusions that you want to take home at the end of the day.

The best advice to give where statistics are involved is to ask questions when things get muddled. Any presenter should be willing to explain their research results if they took the time to include those results in their presentation. Do not be shy or intimidated about statistics, because a little understanding can go a long way for everyone in the room. ■

Chad Follis is a Horticultural Instructor at Mineral Area College in Park Hills, MO. Brad Fresenburg is an Assistant Extension Professor of Turfgrass Sciences at the University of Missouri in Columbia, MO. To see a list of references for this article, see www.sportsturfonline.com