ASK THE "EXPERT"

Luke Cella, CGCS *Tamarack G.C.*

The Language of Statistics

Why do we use statistics? One common and simple reason is to make sense of data. Everyone deals with data to some degree. As a golf course superintendent, I'm always interested in the number of rounds played on my golf course; I'm interested in the air and soil temperatures, the forecast for the following days, the amount remaining in my budget for the month, the number of employees that may call in sick on a given weekend morning. Arguably, the above are only numbers, amounts or degrees. I agree, to the extent that each is only a single amount or measurement; however, when combined with previous amounts or measurements, these individual items collectively become data. It makes sense to look at the whole picture instead of one micro-point when faced with the decision-making process. When we decide to irrigate, we take into account many factors such as air temperature (the days past, current and predicted), soil moisture (the days past, current and predicted), wind speed or velocity (the days past, current and predicted), relative humidity (the days past, current and predicted), etc. When these factors are combined, a data set is formed. From this data set, we draw our decision whether to irrigate or not. We are continually trying to make sense of data in our own way using statistics.

Air temperature is measured in degrees Fahrenheit in this country. So what does this number really represent? It is a numerical representation of how hot or cold the air feels to us. This representation makes it easier for us to discuss the temperature in terms of a number that represents a feeling. It allows us to be more accurate when discussing the temperature around us. Imagine if the only way we could describe the temperature was really cold, somewhat cold, cold, somewhat warm, warm, hot and really hot. In contrast to this method, we use an infinite set of numbers to describe air temperature. In similar fashion, statistics allow us to explain or represent data in a way that is understandable and comprehensible.

It is important to have a basic understanding of statistics for several reasons. We should know how to evaluate a set of numerical facts. Not only is this important in our turf world, it is important in our everyday life. We are exposed to manufacturers' claims for products, to the results of consumer, sociological and political polls, and to the published results of scientific research on a daily basis. If we don't understand the statistics used to describe the numerical facts, the facts become irrelevant.

We should know how to evaluate a set of numerical facts. Not only is this important in our turf world, it is important in our everyday life.

What Do They *Mean*, Anyway?

When people talk about the mean or average, most of the time they are referring to the arithmetic mean of a set of data. The arithmetic mean is defined as the sum of the measurements divided by the total number of measurements. The arithmetic mean is a very useful measurement but it can be distorted due to the presence of extreme values (also known as outliers) in a data set.

Sometimes the average is considered to be the mode. The mode is the most frequent or probable measurement in the data set. The mode is commonly used to measure the popularity that reflects an opinion. It is common to hear the

(continued on page 22)

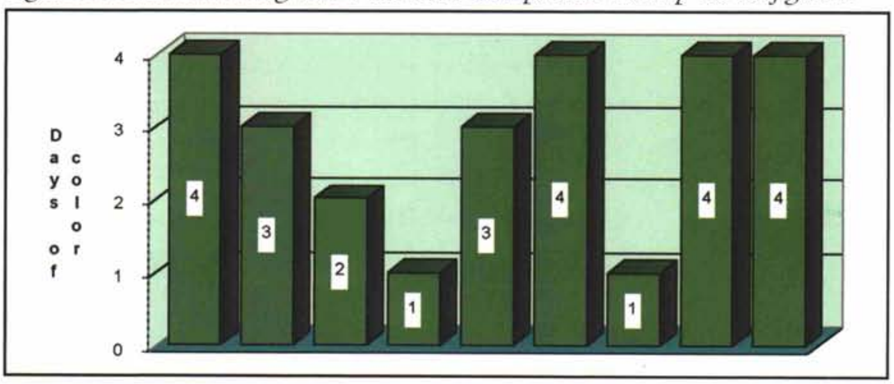
phrase "the most popular model" when discussing some product. This is the mode, or the product that was chosen most often by a consumer group.

In other cases, the average is interpreted to be the median. The median is the central value of a set of measurements; 50% of the measurements fall above the median and 50% fall below it when the measurements are arranged from lowest to highest. The median is most often used in a large set of measurements such as a union wage scale or when dealing with the country's census figures.

The mean, median and mode help to describe what is known as central tendency. Central tendency describes the center of a distribution of measurements and how the measurements vary about the center of that distribution.

Figure 1 represents the author's contrived manufacturers' data on the effects of a new chelated-iron product on turfgrass. Upon application, this product turns the turf a dark-green color. In the graph, each bar represents a golf course and the measured number of days this new chelated-iron product kept the turf dark green. In this data set, the arithmetic mean is 2.8 days ((4 + 3 + 2 + 1 + 3 + 4 + 1 + 4 + 4)/9), the mode is 4 (the number that occurs the most often in the data set), and the median is 3 (in this ranked set of nine numbers [4 4 4 4 3 3 2 1 1], the middle number).

Figure 1. Data describing a new chelated-iron product to keep the turf green.



How many days of color would you expect to get from an application of this new chelated-iron product? Based upon the arithmetic mean, you might expect the color to last 2.8 days. However, if the average is interpreted to be the mode, the color might be portrayed to last up to four days. Measures of central tendency can provide a quick way to characterize a set of data. However, measures of central tendency can be misleading if left undefined. It is imperative to understand what measure of central

(continued on page 24)

tendency is being used to describe the data if the "average" is figured.

Who's Describing the Data?

I am figuring out my capital expenditures when preparing my budget for the next fiscal year. I have been told to budget based upon a "normal year." I look at what has been spent over the last 11 years and this becomes my data set. (Table 1)

Table 1. Capital expense budget for the last 11 years.

Fiscal Year	Capital Expense Budget
1991	10,000
1992	25,000
1993	10,000
1994	25,000
1995	20,000
1996	100,000
1997	10,000
1998	10,000
1999	25,000
2000	10,000
2001	20,000

The arithmetic mean of this data set is $24,090, or $265,000/11 (total capital spent divided by 11 years). I put into my budget the amount of $24,090 (Table 2) for the next fiscal year. After review, the club's general manager changes my figure to $20,000. He uses the median to calculate the capital expense (middle number in the ranked list of expenditures over the last 11 years). After submittal to the green committee, my capital expense budget becomes $10,000. The committee uses the mode (the number that occurs most frequently) to determine the capital budget for the following year. Who is describing the data and what is the intended outcome?

Table 2. Persons representing capital budget data.

Person(s)	Amount budgeted	Method
Superintendent	24,090	Mean
General manager	20,000	Median
Green committee	10,000	Mode

In this example, no one is wrong in the determination of the average spent on capital over the last 11 years. The general manager, the committee and I myself each represented the data in a different way. The data or facts did not change, just the way in which they were interpreted and presented. It is important to know who is presenting the data and the reasoning used to support the facts.

Visual representation of statistical analysis attempts to display information in a simpler form than just pages of numbers.

Visual Representation Is Good—Isn't It?

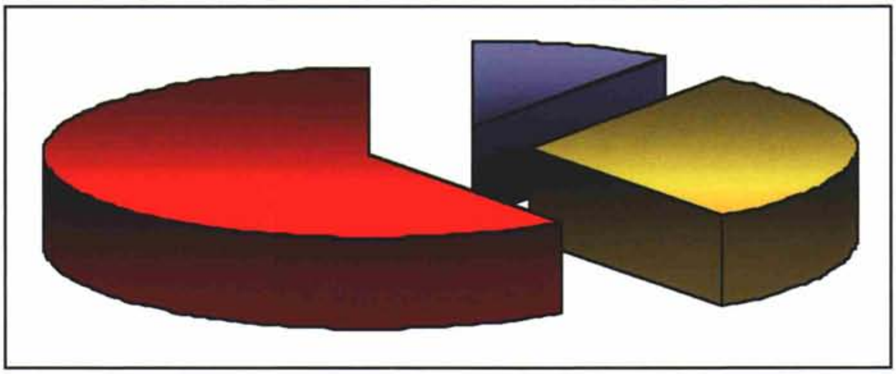
Most statistical analysis of data attempts to make sense of facts. Visual representation of statistical analysis attempts to display information in a simpler form than just pages of numbers. Our media life is full of color. Different visual stimuli through print, television, movies and the Internet continually bombard us. Visual portrayal of statistics comes through the use of tables, charts and graphs. When we look at these visual representations, however, we must take care. Ever take a look at the "before" and "after" shots of someone on an amazing diet? Take a closer look next time. I don't think I have ever seen the person in the "before" shot in a positive light; the subject seldom smiles, he or she has a slumped posture and is never in pinstripes. The "after" shot portrays a happy, smiling person standing in a way that portrays confidence and self-satisfaction. When looking at a table, chart or graph, it is important that we understand the source of the information and the definition of the terms being used to describe the results.

A common method of visually representing statistics is the "pie chart." A circle is divided by a number, usually a percentage, each segment displayed as a wedge or part of the circle. In Figure 2, the pie has three pieces. The largest piece represents 60%, the middle piece represents 30% and the smallest piece represents 10% of a population. The common misconception with pie charts is the size of the wedge. The size of each wedge is a percentage of a complete circle of 360°. The wedge representing the 60% proportion is actually 216° of the circle and to us humans this can be difficult to grasp when we are accustomed to basing percentages on a 0-100 scale.

Another way to misrepresent data is the improper use of the common line graph. These graphs use lines to connect data points between pieces of information. Manipulation of these graphs is commonly characterized by improper or uneven spacing of the scales of the x- and y-axis.

(continued on page 27)

Figure 2. A pie chart representing 60%, 30% and 10% of a population.



Bar graphs use bars to visually depict two or more variables in comparison. Figure 3 is a bar graph and represents two different variables, women and men beginning to play golf in three different years. Each variable (women and men) is exactly equal over the three years; however, the bar graph misrepresents the men. From the depiction, it seems that fewer men are taking up the sport when compared to the women in each of the three years. The bars representing the men are a different shape and size (width) than the bars representing the women in the graph. This gives the illusion that one is less than the other. However, the height of the bars is equal for each group in the same year. Interpreted correctly, no difference exists between the number of men and women taking up the game of golf over the three-year time span. It's always prudent to take notice of areas that are out of proportion to each other, horizontally or vertically, in any graph.

Charts, tables and graphs are a fantastic tool to communicate otherwise impenetrable results. However, taking care to understand the source of the information and the terms being used in the description is vital.

A Percentage of What?

As discussed earlier, a percentage is a very common way to represent some type of information. In the case of the pie chart (Figure 3), three percentages were given: 60, 30 and 10. In this case, they add up to 100%, and in a perfect world (such as creating examples during the writing of an article) 100% can be reached easily. In the world of data, though, 100% is often difficult to achieve. Often missing data, processing errors and/or sampling errors preclude arriving at 100%.

It is important to know the basis of the percentage reported. A situation where we may encounter this could be on a seed tag (Table 3).

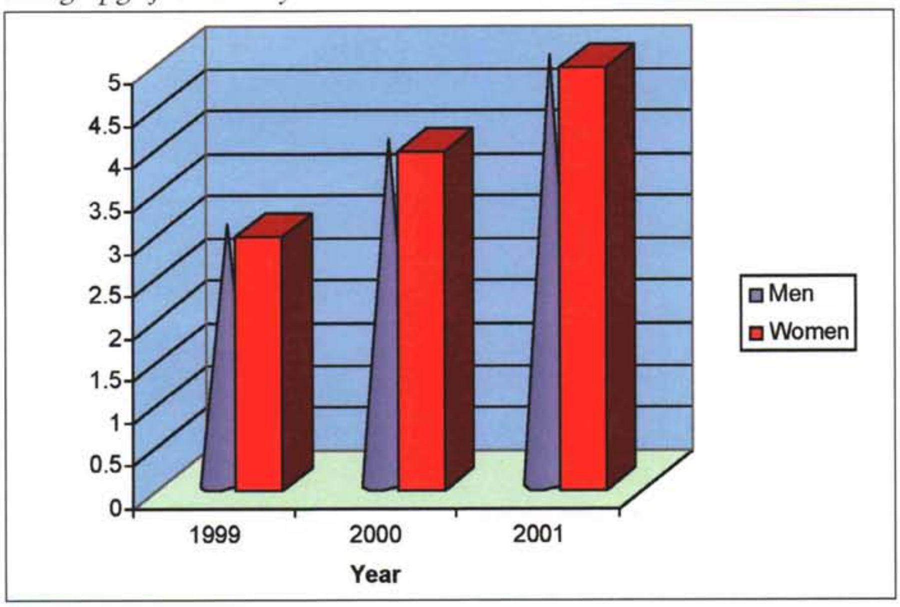
Table 3. The seed tag.

Authentic Kentucky Bluegrass Seed	
Pure Seed	**97.89%**
Other Crop	**0.99%**
Inert Matter	**1.12%**
Weed Seed	**0.00%**
Total	**100.00%**

This seed seems to be pretty pure with 97.89% of the seed being pure seed, 0.00% weed seed, and 1.12% inert matter. However, the "other crop" component needs some attention. This part is the total percentage of mature, sound seeds of other crops found in the representative sample tested. These may or may not be objectionable seeds depending on your use for the primary kind of seed.

The seed tagged is Kentucky bluegrass and this is a 50-lb. bag. Of that 50 lb., 48.94 lb. is pure Kentucky bluegrass seed. If 0.99% of those 50 lb. is seed of another crop, there is roughly 0.5 lb. of that other crop. If that other crop happens to be creeping bentgrass (and it is another crop), and in one pound of creeping bentgrass the mean number of seeds is 6,412,000, there could be up to 3,206,000 bentgrass seeds in that bag of Kentucky bluegrass. On closer inspection, the 0.99% other crop found in this bag takes on some importance (to the tune of 3,206,000). In this instance as in all, it is important to know the base of your percentages.

Figure 3. The bar graph representing women and men taking up golf over three years.



Don't Fear the Analysis

Mark Twain can be credited with the line, "Facts are stubborn, but statistics are more pliable." This truism reveals how necessary it is to have an understanding of statistics. We live in a society where we are pummeled with statistics on a daily basis. We are drawn to statistical outcomes, polling and indices because they appear interesting. We make judgments and decisions based upon statistical outcomes in our personal and professional lives. Often we take the time to read the statistics on a product we may purchase or a method we may try, but seldom do we take the time to read the statistics on the statistics.