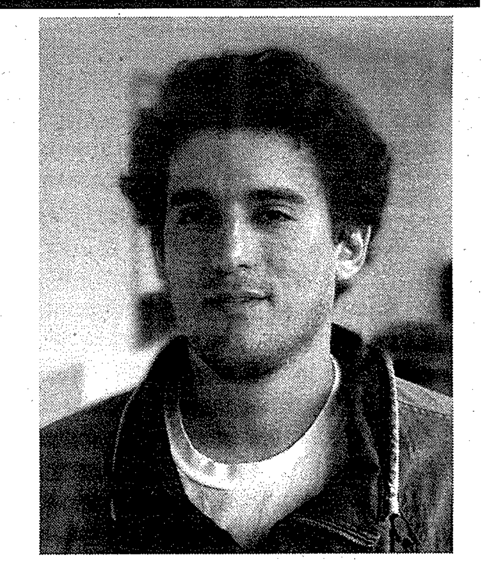
ENVIRONMENTALLY SPEAKING

Peter Schwartzman

Statistics 101



Everyone needs to grapple with stats

Statistics are everywhere. And their power of influence is immense. Unfortunately, it is very easy to get confused by them. But given how important and powerful they are, we, acting as informed citizens, must work extra hard to comprehend them and understand how they can be used. In what follows, I plan to introduce and tackle a few of the more common and vexing statistical concepts. I will also attempt to illustrate how they might be used in an environmental context as well. So, sit back, put your feet up, and let's get "statted."

Mean versus Median. Almost any scientifically-related report uses the words "average" or "mean." These mean the same thing; pardon the pun. The average of a variable (say, temperature) is just the sum of all its values divided by the number of such values. So, for instance, let's say the daily temperatures for a period of one week are (in Fahrenheit): 65, 70, 60, 65, 70, 50, & 60. The average (or mean) temperature for the week is 62.9°F. Now, the median temperature for the week is 65°F which is the "middle" value of the dataset (when they are put in numerical order). Thus, here, the mean and median are not the same. Generally, it is common to have means and medians that are different from one another. And this difference can lead people to focus on one value, rather than the other, depending on their underlying goals. In the case presented, if someone were trying to demonstrate "global warming" they might talk about the median, or if they were trying to suggest "global cooling" they might reference the mean/average (since it is "colder").

But readers shouldn't think that the choice between them is usually done for manipulative purposes. The median, which is less commonly used, may be a more useful statistic than the mean in some instances. For example, if one wants to look at the number of children that people are having, it might make more sense to know what the median is rather than the average. This is because the median will be an integer—1,2,3, etc.—whereas the average is likely to be a number with a decimal component. And since people aren't having half-children, it may make more sense to look at the median value.

Correlation versus Causation. When we want to know the strength of a relationship between two variables, we can use a simple (relatively speaking, of course) calculation to assess this; we calculate a correlation (technically, a correlation coefficient). A "strong" relationship (represented by a large correlation) indicates that the two variables tend to "follow" each other—that is, when one changes the other changes as well (typically in a linear fashion). If the variables behave similarly (that is, increase together *or* decrease together), we say that they are positively correlated and if they behave "inversely" (that is, one increases while the other decreases), we say that they are negatively correlated.

So for instance, we might think that warm days occur when it is sunny. Therefore, we would expect a calculation of the correlation between the daily maximum temperature (written, T-max) and amount of daily sunshine to be high. However, it turns out that these two variables are positively correlated in summertime, when the sunny days tend to be the warmest, but are often actually somewhat negatively correlated on winter days when some of the coldest days are actually quite sunny. Thus, an examination of the relationship between the amount of sunshine and T-max would only show a high level of correlation if we looked at only one season at a time and wouldn't show much correlation at all if we looked at all days of the year.

In the field of environmental studies, researchers are often asking if two variables are related. And often the first step taken is to calculate the level of correlation between them. But just because two variables are correlated doesn't mean they are causally related (meaning, a change of one variable **causes** another variable to change). For example, many people think that poor people have more children than wealthier ones. On a global scale, there definitely does appear to be a relationship (i.e., correlation) between the relative wealth of a country and the average family size found among its people. However, it would be misleading (and, perhaps, flatly wrong) to say that, based on a confirmed high correlation between these two variables, poverty causes families to be large (or that large families lead to conditions of poverty). While this conclusion may be true, it isn't determined by a correlation calculation alone. This is one way correlation can be misused.

Since causation ("what causes something") cannot be easily ascertained in many cases, especially environmental ones (given the extremely large number of contributing factors), it is often best to reserve judgment on "cause" and adopt a different framework of understanding the problem. In order to determine the "cause" of something, we often have to conduct a much more complex set of analytical and statistical procedures. Even then "proving" causation "beyond the shadow of a doubt" is very difficult. Let's look at just one example that illustrates why this might be.

Let's say that Chemical J has been shown to cause cancer in mice. In fact, it has been found to cause bladder cancer much more frequently than other forms of cancer. Now, let's say that a person, named Joe is diagnosed with bladder cancer. Did Chemical J cause Joe's cancer? Obviously not, right? Well, actually, we don't know until we attempt to find out. First, we may want to know if Joe had any exposure to Chemical J. Let's say he did when working in a factory. But just because he was exposed doesn't mean that he got cancer from it. What do we ask next? There is a long list of questions that we might ask, including: Have other employees in his workplace gotten bladder cancer? What other chemicals (or lifestyle choices, etc.) are associated with bladder cancer? Was Joe exposed to these? Are there groups of chemicals that work together to enhance each other's carcinogenicity? (Though is not uncommon to find chemicals working together—synergistically—to cause something to happen, it is very difficult to conduct research that looks at these relationships.) Are some people genetically predisposed to bladder cancer? Does Joe have these genetic traits? And yet, even after we get answers to all these questions, we cannot be sure that Joe got bladder cancer from Chemical J. This is because there are so many confounding factors in the development of cancer and cancer often doesn't express itself for ten to thirty years after the relevant exposure. But since it is so important that we know why he got bladder cancer (as part of a legal response or to prevent others from getting it), it is stifling to realize that it is nearly impossible to know if chemical J caused Joe's cancer. So where does this leave us?

Since so many environmental problems (such as, climate change, cancer clusters, respiratory disorders, species extinction, etc.) have serious impacts on us, we have a great desire to know why these problems arise. As shown in the previous example, our ability to determine the "why" (a question of causality) can be quite complicated or near impossible. This realization can lead us in one of two directions. Either we can, as we largely do now, put decisions through a cost-benefit analysis which requires the assignment of a monetary value for a human life and a weighing of a plethora of inherently uncertain statistical findings. Or, it might lead us to live more humbly and avoid complicating the environment excessively/needlessly (through the use of synthetic pesticides, for instance) because we don't really know what the overall impacts will be. This latter position, to which I generally subscribe, doesn't mean that we must live in a "box" and avoid all exposures to risk but rather it means that we don't allow the use and emission of potentially dangerous chemicals precisely because we are aware of the inherent uncertainty that exists in attempts to assess their impact and establish causality.

Correlation between variables shouldn't be ignored, however. Nor should attempts to establish causality be eliminated either. In fact, both correlation and causation have their place. They can be used to identify relationships (even potential or suspected ones) between variables. Ultimately, however, readers of statistical results (as found in any daily newspaper) should reserve judgment and maintain a skeptical position. Whether one is more skeptical about reports of environmental catastrophe or claims about the amazing resilience of ecosystems is largely a personal decision, but it is one that we should all recognize is predicated on the certainty (or the lack thereof) that we attribute to the power of statistical analysis and the complexity of living systems.

Significance versus Importance. Probably one of the most commonly used words involving the reporting of statistics in the mainstream is "significant." What does it mean to say a result is significant? Does it mean that it is "important" or does it mean something else?

Within the world of statistics, when the word "significant" (or "significance") is used, it means that a result is "not by chance" and, therefore, represents a meaningful finding. Unfortunately though, this term gets thrown around so much it is very difficult to know what it means. For a clear demonstration of the confusion that can arise, consider the following example. Statement 1 reads: "The increase in global temperatures is significant." Statement 2 reads: "There is a significant increase in global temperatures." Statement 1 speaks to the "not by chance" character of the observation—here, the increase in global temperatures. Statement 2, on the other hand, says something about the importance of the result. And, no, these are not the same thing. Let's understand why. If a finding is "significant" (ala Statement 1) it means that it is a result (here a relationship between two variables—time and temperature) that is not expected to be found in random data more than a certain amount of the time (usually, less than 5%). So if the result isn't random it means that there **is** a relationship (here, a trend in temperature) that is real. If a result is said to be significant in extent (ala Statement 2) then it means that we should be alert to its potential impact on something of importance (here, rising temperatures of a certain level may melt ice sheets and cause sea level to rise appreciably). Another example helps illustrate the distinction further.

Global population increased quite a lot during the 20th Century. Currently it is rising at just above 1% per year. An analysis of the actual data from 1950-2005 reveals some interesting findings. The correlation between population and time (over this period) is 0.998—an extremely high value (the maximum being 1.00). (Interesting, the correlation between the population growth rate—PGR, in %—and time is -0.68%, indicative of the rather steady decline in PRG since the early 1960s.) One might (falsely) conclude from this result that time is "causing" populations to grow. And while it certainly takes time for populations to grow, it is people procreating that is "causing" populations to grow. And, arguably, populations are growing (rather than steady) because economic and political forces are motivating families to reproduce more than replacement level (which is about 2.2 children per family).

Once we suspect a relationship (based on a high correlation value), we then perform a linear "regression" analysis to determine if the trend is statistically-significant. Simply, this means that we attempt to determine if a there exists a change in human population size that we can say is "not expected given random data." Such an analysis here establishes that a trend does indeed exist (i.e., it is "statistically-significant"). But even if a 1.68% annual growth in population (which is the average growth over the fifty-five year period) is statistically-significant, might it be more relevant to know if this trend is important? In other words, will this trend create problems for societies in the future? Doesn't our answer to this depend on several other pieces of data, such as, "How big is the population to begin with?," "Are future people going to drive Hummers or are they going to use public transportation?," and, "Will oil or coal be our future energy source or will the transition to a renewable energy economy occur quickly?" Depending on how we answer these secondary questions will obviously impact our evaluation of the importance of a 1.68% growth rate. In some cases it will be very important. In others it may be much less important.

From now on, I hope you will never be tricked into thinking a median is a mean or significant findings are (necessarily) important. Statistics can be misused intentionally (as well as unintentionally) and therefore it is imperative that we all become more familiar with this "foreign" language. Given the ubiquity of their existence in our lives, stats compel us to avoid becoming mere statistics.

Peter Schwartzman (email: drearth1@gmail.com) is associate professor and chair of the Environmental Studies Program at Knox College. Father to two amazing girls, Peter hopes that their lives will be lived on a cleaner, more just, planet. A nationally-ranked Scrabble® junkie, he is also the founder and maintainer of websites dedicated to peace, empowerment, and environmental well-being: www.onehuman.org; www.blackthornhill.org; & www.chicagocleanpower.org.