

How Statisticians Think and Why It Matters

Drug Information Association
Thirty-seventh Annual Meeting
Denver, Colorado
10 July 2001

Rafe M. J. Donahue, Ph.D.
Principal Consultant Data Mining
GlaxoSmithKline
Research Triangle Park, North Carolina
rd34115@gsk.com
<http://home.earthlink.net/~rafedonahue>

Some background

“I’m not a statistician but I play one on TV.” That’s how I often tell people what I do, using a reference to a commercial from some years back in which an actor, who played a doctor on a daytime television program, used his pseudo-authority to hawk some pain killer. Although that response typically produces a chuckle, it often also produces a response thinly related to statistics (“Golly, what are the odds of that?”) but more often produces a comment exposing some loathing or fear of the field of statistics (“Wow! I hated that class in college!”)

I’ve been doing statistics for about a dozen years, professionally for almost ten. I have a science and math background, with more hours than I care to count in biology, physics, chemistry, and the like. Like most statisticians, I’ve seen lots of good and bad statistics and lots of good and bad science. And I’ve also seen the statistician often on the other side of the fence, when it comes to styles of thinking, from his or her scientific colleagues, particularly, due to the nature of my employer, physician and clinicians.

What drives this difference in thinking styles? From where does it come? Are statisticians responsible for producing it? Is there something in our training, or in the science training of our colleagues, that produces these different mindsets?

I decided to examine some fundamental concepts of statistics and how I look at these concepts in my work compared with how my scientific colleagues view them in theirs. Thus, this work-in-progress represents what I have discovered thus far.

Four elementary differences

I found four elementary areas with differences between “traditional” and “statistical” thinking styles. These areas are shown in Table 1.

Table 1. Four differences in thinking styles

	<i>Traditional thinking</i>	<i>Statistical thinking</i>
Event reproducibility	Certainty and determinism	Randomness and stochasticism
Acuity	Individuals	Distributions and populations
Precision in determination	Certain knowledge	Estimation bias and variability
Granularity	Look at all the data	Data reduction

These areas are not mutually exclusive, nor are they exhaustive. They simply represent arenas in which statisticians need to assist our scientific colleagues in seeing the other side of the coin. Each of these areas is discussed below.

Reproducibility of events

I was taught in my science courses that scientific investigation demands repeatability and that our physical world behaves according to certain deterministic rules. The most notable example of this deterministic thinking comes from my physics classes. I was taught that the distance d a body moves in a time t when under a constant acceleration a is described by $d = \frac{1}{2}at^2$.

I remember using this formula, and some others attributable to Newton, in a little experiment in high school. We had small steel ball, a marble, of a certain mass. We were also given a steel ramp shaped like

a “j” when viewed from the side but its cross section was a “u” shape just narrower than the width of the ball. This ramp could be secured to one of those science benches that existed in all the science rooms and could be used to launch the ball horizontally. That is, we would put the ball at the top of the ramp and it would roll down the rails of the ramp, picking up speed, and get launched horizontally across the room.

The point of the whole experiment was to convert potential energy to kinetic energy and then compute, using our fancy physics formulas, where the ball would land. We did all sorts of measuring and weighing and computing and determined, eventually, exactly how far from the edge of the table the ball would sail after it was launched from our ramp.

At the appropriate spot on the floor we taped a piece of the old-fashioned carbon-copy paper, the stuff that leaves a mark when struck. On the paper we drew a line exactly where the ball was supposed to land. (And, being the goofs that we were, we also drew a little military installation — tents, jeeps, latrines, the whole nine yards — on the paper, in an effort to add a little fun to whole endeavor.)

So, we get it all set up and let 'er fly — and the ball doesn't even come close to hitting our little Army camp.

Someone decides that we didn't release the ball cleanly and we need to try it again. And we miss again.

Someone else decides that the other person didn't do it right so we switch launch personnel. And we miss again.

(At this point, one has to imagine the chaotic scene in the classroom with a dozen groups of students launching marbles off of desks and chasing them caroming off chairs and desks and other students. It must have been a zoo.)

We, of course, do the experiment over and over again, changing this and that and whatever, and we check our math over and over again. Eventually we cheat and pick up the tape and move the camp and are able, with great ballistic fanfare and exclamation, to take out the motor pool. But subsequent iterations yield puzzling results. Even after moving the paper, we don't always hit the motor pool. And we never do hit the line.

There was tremendous variability that was out of our control hiding in our process. We weren't witnessing $d = \frac{1}{2}at^2$; we were witnessing $d = \frac{1}{2}at^2 + e$, with e being some sort of random error in the process. This error wasn't in the textbook formulas; what was the deal?

The deal was that the same process doesn't always produce the same result. Things aren't deterministic as the formulas would have us believe; they are *stochastic*, fraught with variability that is beyond our control and often beyond our understanding.

One could argue, of course, that the reason we got different results each time is that we actually did different things each time. Once we released the ball slightly higher than the time before. The ball got warmed in our fingers and rolled with a different amount of friction. The air conditioners moved the air more one time than the other.

All of this is true, of course. But just stating those facts doesn't help us do science. You see, when we do the experiment, we need to deal with this variation and distribution; we can't just say that it is a different thing each time. If it really is a different thing each time, then there can be no such thing as reproducibility and science is doomed. Realization of the natural, uncontrollable variability and use of statistical approaches to deal with that variability actually liberates us to find and identify the processes underlying the observations that we make.

Acuity in viewing experimental units

Examine the game of golf. Golfers carry around a bag of clubs that have different properties. The clubs with low numbers are used to hit the ball a long distance with less height while the clubs with high numbers are used to hit the ball a shorter distance but with greater height. Part of playing well, I am told, is knowing which club to use when. So when you walk up to the ball lying in the fairway (sometimes it lands there, really), you must judge the distance to the hole and then which club to use. The golfer who is really good will know exactly what club to use if he or she is very consistent. That is, the good golfer will know that, say, her 5-iron will hit the ball 160 yards, give or take a few yards. (A golfer like me knows that my 5-iron will hit the ball about 160 yards give or take about 100 yards!) What is of value here is to note that I don't get the same result each time I swing. The distances have a *distribution*. If I have knowledge of that distribution, I can make wiser decisions about club selection. If the hole is 160 yards away and there is water behind, I probably need a shorter club. If the water is in front, I need a longer club. (If I were more consistent it might not make a difference.) Since I'm aware that the distance the ball will travel comes from a distribution, and I know some things about the distribution, I can make adjustments to improve the

chance that I will be successful. My focus cannot be on the distance produced by a single swing but must be on the distribution of such distances. I cannot afford to focus on *individuals*; I must focus on the *distribution of the population*.

In the world of clinical trials, the measurements we make (airflows, blood pressures, depression scales, yes/no responses, numbers of seizures, etc.) have distributions because they come from different patients or from the same patients at different times. So we need to think about what the distribution of patients is doing, not what the individual patient is doing, to think statistically and scientifically. This seems to fly in the face of clinical practice (“Each patient is unique and must be treated as such.”) but it is the only way in which we can do good science. If all the patients are so unique, why do we collect the same vital signs on each of them? Why the same general questions in a patient history? Why compare patients to the same growth charts? The answer is that, although the patients are all unique and require special care, all we can do to determine what needs to be done is to compare the same measurements on each patient to the collective norm. That is, we compare apples to apples and we have knowledge of what the *distribution* of these measurements is when viewed for healthy patients. Only if the measurements in question are inconsistent with the distributions in healthy patients can we determine that the patient is unhealthy.

This, of course, does not mean that the clinical practice of treating patients should become simply a cold, impersonal, look-it-up-in-some-big-book kind of endeavor. Medical practice has always had certain aspects that make it a mixture of science with art. I do not intend to debate the art portion of medicine; I am here to address the needs of the science portion. And to do so requires that we think statistically. And to do that means we need to think about distributions, not necessarily about individuals.

Precision in making determinations from data

Scientific investigations typically involve sampling individual units from some grand population of units. This process of selection induces error into estimation since, as we have seen above, the same process (selecting a sample) doesn’t always produce the same result. Thus, any computations that one makes on sample data that are then used to make determinations must be subject to this sampling variability. As such, any estimate of a population characteristic carries with it an inherent amount of uncertainty.

We must always attempt to measure and describe this uncertainty; we cannot treat our estimates as certain knowledge.

Thus, when we report proportion estimates, we cannot only say that 31.6% of subjects treated with drug X showed signs of improvement but we must include some measure of the uncertainty of the estimate; we must include a margin of error, as the pollsters do at election time. To state an estimate as a certain truth when it is based on a sample subject to sampling error is misleading and constitutes scientific corruption. Really.

Furthermore, estimates often possess bias, a systematic error that distorts the estimate’s ability to represent the truth. That is, an estimate based on data always estimates something; it just might not be the thing that you think it is or want it to be. The good scientist should always look for sources of bias and present such issues to allow the reader to interpret the effect. The ways we select patients, design questionnaires, and take measurements all can influence bias.

[And a point about *false* precision, a pet peeve of mine: Flight times are often given to the nearest minute. Do they really think that we think the plane will leave exactly at 4:41? Why not just round to the closest 5 minutes? Unemployment rates are rounded to the nearest one-tenth of one percent, an effort, I heard someone say, to demonstrate that the people in the Labor offices have a sense of humor. Do they really think that their estimates are that precise?]

Granularity of examining data

Picture a typical clinical trial: two treatment groups, multiple visits for each subject, multiple types of data collected (adverse events, demographics, concomitant medications and illnesses, dosing compliance, efficacy measures, vital signs, etc.). Perhaps hundreds or thousands of data points are collected on each subject. There are dozens or hundreds of subjects in each treatment group.

Then, at the end of the trial, the data are “analyzed”. The primary endpoint is examined and the trial is deemed a success or failure. From the hundreds of readings on each subject, a single efficacy summary value is computed (change from baseline, time to some event of interest, response or nonresponse). These summary measures within a subject are combined across subjects in the same treatment group to produce some summary of that treatment group (average change, median time to event, percent of responders). The

two groups' data are then combined into the one magical tell-all summary of the trial: the p value. All those data points are distilled down to one number upon which the whole trial rests. That's data reduction.

It is important to realize that there are countless ways to reduce the data. Repeated-measures studies can reduce multiple observations per subject down to a single summary through the mean or the median or the maximum or the minimum or the area-under-the-curve or the slope of the regression line or the number of extreme values or whatever. Each of these summary measures within a subject tells a different story about that subject.

Furthermore, reducing the data across subjects within a treatment group can also be done in a variety of ways: mean, median, etc.

And still further, there are different ways to compare the treatment groups: difference between groups, odds ratio, hazard ratio, overall F tests, contrasts, etc.

The point here is this: in this grand and wonderful world of science and mathematics there is only one group of things that can be *compared* and that set of things is *scalars*.

Scalars, as opposed to vectors, are just numbers. We can compare numbers: 3 is greater than 2, 7 is less than 31, 5 is equal to 5. Vectors are ordered collections of numbers: (5, 9, -1, 2), or (0, 1, -8), or (6, -4, -9, 0, 7, 4, 4, 17) but we cannot compare these collections. What is bigger, (2, 5) or (1, 9)? We can compute the magnitude of each vector and compare them but magnitudes are just scalars. We cannot compare the vectors themselves.

Suppose my daughter asks me whether a tree is bigger than a train. How do I answer that question? I can take a number of measurements of each (height, weight, length, girth, mass, volume, surface area, etc.) and put them in a vector of readings but I still can't compare them until I reduce the data vectors down to scalars. The reduction might be something simple, like just pulling off the height component, or it could be more complex, like computing the surface area to volume ratio.

Subjects in a clinical trial generate data vectors. The analysis process reduces the vectors down to scalars that can be compared. And here is the important thing: the way through which one reduces the data can make or break a study. The right reduction can demonstrate a treatment difference; the wrong reduction can cloud a difference that exists.

Science and statistics require comparisons. Comparisons necessitate scalars. Data reduction is the process by which vectors are refined into scalars.

Now, the temptation exists to avoid data reduction. Some people will claim that they just need to "see all the data" in order to make a conclusion. While I agree that there is value in the joint distribution of some of the responses, attempting to look at "all the data" is a pipe dream.

We recently ran a study with about 40 exploratory, secondary endpoints. Two analysis populations were defined, ITT and protocol correct. Two strategies for dealing with dropouts were under consideration, use only the observed data and use last-observation-carried-forward imputation. Furthermore, adjustments were made for change from baseline and percent change from baseline in addition to the 'raw', unadjusted data. And each measure was determined for each subject at something like eight posttreatment visits. And the clinicians were interested in at least 8 different subgroup splits of the subjects, in addition to the collection of all the study subjects.

Doing the math, one can conclude that there are in excess of 3800 combinations being examined here, not counting the multiple visits! I asked the clinicians to pick a "most important" scenario for analysis, knowing that there would be more tables than I wanted to produce. "Just run all of them," they said. "We really need to look at all the data to be able to understand what is going on in the study." Yeah, right.

Judicious data reduction is essentially to develop understanding.

Why does it matter?

The natural processes we observe are frighteningly complex. Factors that we understand, factors that we don't understand, and even factors that we understand but cannot predict (stochastic ones) are at work in generating the data that we see. Failure to deal with all the factors that are behind the scenes prohibits us from successfully understanding what nature is trying to tell us. To maintain our scientific integrity, we need to be humble enough to deal with our data from a statistical point of view: to see that some things need to be viewed as stochastic, that understanding of the distribution is key, that estimates are subject to variability and bias, and that data reduction is necessary to be able to grasp complex and large amounts of data.