Geoffrey Loftus: Personal Reflections, April 10, 2010

I am, by nature, sufficiently dimwitted that I can only understand some concept when it's presented to me in an extremely clear and straightforward manner. So I felt pretty confused when I first learned about statistical hypothesis testing as a Brown University undergraduate in 1966, and felt only marginally less confused when I learned about it again as a Stanford University graduate student in 1969. To me, hypothesis testing seemed complicated, unintuitive, misleading, and, in general, a profoundly indirect way of trying to unveil what a data set was trying to tell you.

So was there a reasonable alternative to deal with the inevitable statistical error that accompanies data collection? Yes! I began, in my early years, to figure out that the process of plotting data accompanied by confidence intervals seemed to accomplish the same thing as hypothesis testing but in a manner that was clearer, more intuitive, and more straightforward. I noticed, meanwhile, that other sciences—Physics, say—made zero use of hypothesis testing as a means of transiting from data to conclusions, while at the same time making considerable use of confidence intervals. To me this constituted an important clue about why there seemed to be a lot of progress in those sciences, compared to a lot of confusion, misunderstanding, and backfilling in Psychology.

I didn't do much about this personal angst beyond whining to my colleagues and stubbornly refusing to carry out hypothesis testing in my own research articles until 1991 when I was asked by *Contemporary Psychology*, to review a marvelous book about the history of statistics called *The Empire of Chance*. I entitled my *CP* review "On the tyranny of hypothesis testing in the social sciences" and used the review (as is so subtly suggested by its title) as a vehicle to articulate all that that I found lacking in current statistical practice within Psychology.

Over the intervening two decades, I've taken a more active role, issuing forth talks, articles, and chapters bemoaning the prevalence of hypothesis testing and suggesting alternatives, particularly the use of data plots and confidence intervals. During the mid-1990's, I agreed to serve a four-year term as editor of *Memory and Cognition* primarily so that I could try to directly influence the nature of data analysis in at least one major psychological journal—to slash the reliance on hypothesis testing while ramping up the reliance on plotting data with accompanying confidence intervals (this attempt met with mixed results). In 1994 I published, with Mike Masson, an article in which we described a new use of confidence intervals in within-subjects designs. I became increasingly strident in crafting my refusals to requests from editors that I include hypothesis testing in manuscripts that I'd submitted to their journals.

I could go on and on here, but there's not space for it. My nutshell summary of my general feelings is, I think, best captured in a soliloquy which I always deliver at some point to the undergraduates in an advanced statistics course that I teach yearly. Roughly speaking, it goes like this.

> OK, everyone listen up. For the next few minutes, what I'm going to say is really important. So please stop texting, web surfing, newspaper reading, or anything else besides listening to me. OK? Here we go.
>
> In the kinds of experiments that we do in Psychology, we would ideally like to find out the values of *population parameters*—typically, although not always, we would like to determine, as best we can, the *pattern of population means* over the conditions in our experiment. We can't determine exactly what these population parameters are because all experiments are bedeviled by statistical error which obscures them. Statistical analysis is largely designed to address this problem in one way or another.
>
> The most prevalent such analysis technique is *hypothesis testing,* whose main goal is to conclude, if possible, that a set of population means does *not* conform to one specific pattern, namely, "they're all the same." If we make this conclusion, that is if we "reject the null hypothesis" we don't get very far because the hypothesis testing process doesn't readily provide us with any idea about which of the infinite set of possible alternatives to "they're all the same" is the correct one. If on the other hand we *don't* make this conclusion, then we're sorely tempted to make the error

of concluding, either implicitly or explicitly, that the population means *are* all the same. In short, hypothesis testing at best provides us (sort of) with one instance of what the pattern of population means that we're seeking *isn't*, and at worst just leads us astray.

A second analysis technique is to plot your sample means with associated confidence intervals around them. This technique provides you, immediately, intuitively, and directly, with two critical pieces of information. First the pattern of sample means that you observe constitutes your best estimate of the corresponding pattern of population means which is exactly what you're seeking. Second, the confidence intervals provide you with a sense of how seriously you should take this pattern of sample means as an estimate of the underlying pattern of population means. Small confidence intervals: take it seriously. Large confidence intervals: don't take it so seriously and go out to collect some more data.

Notice that plotting the means with the confidence intervals tells you pretty much everything you'd find out by carrying out a hypothesis test. The reverse, however, doesn't hold.

So it's up to you to choose between these techniques (or invent others if you feel creative). Any questions? OK, go back to whatever you were doing.