Alternatives to some statistical conventions, Peter J. Taylor, 20 August 2011

The alternatives are offered here in the spirit of critical thinking, namely, that we understand ideas better by holding them in tension with alternatives even if, in the end, we stick with the conventional approaches.

1. Visual exploration of data: The plot below allows risk to be displayed in relation to two variables. The contours on this plot are derived from a non-linear model fit to the data. What we see is that there are regions where slightly below average growth in the first year results in higher mortality than lower growth, even though the general trend is for higher birth weights and higher growth rates to result in lower mortality. This effect would be hard to detect in a conventional analysis of statistical significance.

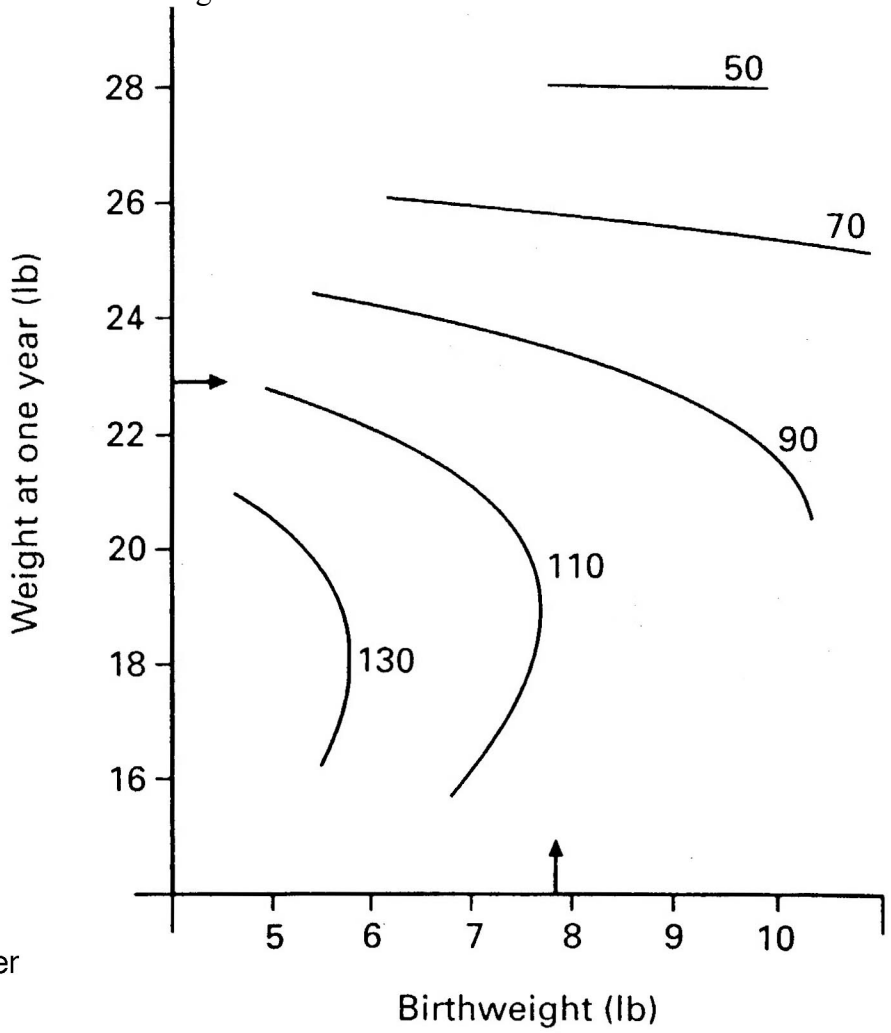

FIG 3.6—*Relative risks for coronary heart disease in men according to birthweight and weight at one year. Lines join points with equal risk. Arrows = mean weights.*

Source: Barker, D. J. P. (1998). Mothers, Babies, and Health in Later Life. Edinburgh, Churchill Livingstone.

Clive Osmond, Barker's biostatistician told me that they stopped using these contour plots soon after the 1989 paper from which this 1998 version is adapted. Because it was not a conventional plot, people said they didn't know how to interpret it or they wanted to see the raw data represented in the plot. Researchers who, in contrast, are interested to do more exploration of data visually may find the following reference illuminating and moderately accessible:

Cook, D. and D. F. Swayne (2007). Interactive and Dynamic Graphics for Data Analysis. New York, Springer.

2. Continuous versus categorical variables: Analysis of categorical variables is the norm in epidemiology, where people are classified as having the disease or not (or dead or not) and many of the conventional statistical methods are built around this. It is sometimes even the case that continuous variables are transformed into categorical so that logistic regression and related methods can be used. Obviously, information is lost in such a transformation and more insight (and higher R^2) could be achieved if the analysis used the continuous variables. Studies referred to in a recent article in the NY Times (August 1, 2011), "Really? The Claim: A Normal Heart Rate Is 60 to 100 Beats a Minute," gave this point a clinical significance. Instead of considering 60 to 100 beats per minute normal and focusing attention on patients outside that range, the studies indicate that "for each rising increment of 10 heart beats per minute, the risk of dying of a heart attack increased 18 percent among women and about 10 percent in men."

3. Correlation, Regression, and Prediction: Everyone knows that correlation is not causation, but most of us interpret regressions in a causal spirit.

From Taylor (2008), http://bit.ly/osTjQ3:
Consider the concept of a regression line as a best predictor line. To predict one measurement from another is to hint at, or to invite, causal interpretation. Granted, if we have the additional information that the second measurement follows the first in time—as is the case for offspring and parental traits—a causal interpretation in the opposite direction is ruled out. But there is nothing about the association between correlated variables, whether temporally ordered or not, that requires it to be assessed in terms of how well the first *predicts* the second (let alone whether the predictions provide insight about the causal process). After all—although this is rarely made clear to statistics students—the correlation is not only the slope of the regression line when the two measurements are scaled to have equal spread, but it also measures how tightly the cloud of points is packed around the line of slope 1 (or slope -1 for a negative correlation). Technically, when both measurements are scaled to have a standard deviation of 1, the average of the squared perpendicular distance from the points to the line of slope 1 or -1 is equal to 1 minus the absolute value of the correlation (Weldon 2000). This means that the larger the correlation, the tighter the packing. This tightness-of-packing view of correlation affords no priority to one measurement over the other. Whereas the typical emphasis in statistical analysis on prediction often fosters causal thinking, a non-directional view of correlation reminds us that additional knowledge always has to be brought in if the patterns in data are used to support causal claims or hypotheses.

[Postscript: The tightness of packing view of regression for continuous variables can be extended to multivariate associations through Principal Component Analysis, factor analysis, etc. The difficulty of interpreting principal components or the factors can be flipped on its head: What causal assumptions about *independent* variables (i.e., independently modifiable variables) enter into interpretations of conventional regression analysis?]

Taylor, P. J. (2008). "Why was Galton so concerned about "regression to the mean"?—A contribution to interpreting and changing science and society." DataCrítica **2**(2): 3-22.

Weldon, K. L. (2000), "A Simplified Introduction to Correlation and Regression," *Journal of Statistics Education*, 8, http://www.amstat.org/publications/jse/secure/v8n3/weldon.cfm, viewed 22 Jun '09.

4. The possible heterogeneity in factors underlying the development of a trait.
When people invoke twin studies or concordance rates to claim that a trait is substantially genetic, they are saying something quite different from when epidemiologists or social scientists find a statistically significant association of some variable with the trait. This difference is often not clear in the discussion of quantitative genetics (QG) research. Contributing to the unclarity are the conventional terminology and lack of attention to the possibility of underlying heterogeneity.

From Taylor (2011)
In this article "factor" is used in a non-technical sense simply to refer to something whose presence or absence can, at least in principle, be observed or whose level can be measured. In genetics a genotype is the set of genetic factors an individual possesses (or the subset held to be related to some given trait). In classical QG, however, the label "genotype" is applied to groups of individuals that are genetically identical (pure lines) or whose mix of genetic factors can be replicated (such as an open pollinated plant variety), or to groups whose relatedness by genealogy can be characterized (such as human twins). No knowledge of actual genotypes is entailed in the QG use of the term. Similarly, the label "environment" is applied in classical QG to the situations or places in which the genotypes are raised without knowledge of the relevant environmental factors…

[T]he intention is to counter any conceptual slippage from analysis of observations of a given *trait* to claims about "genetic" and "environmental" differences. Such claims suggest misleadingly that classical QG analyses of variation in traits address the measurable genetic and environmental *factors* involved in the development of the trait. (For a similar reason, phenotype is not used here to refer to the traits.)…

A corollary of keeping traits and underlying measurable factors distinct is that the factors *underlying* the development of observed traits may be *heterogeneous*, that is, they do have to be the same from one set of relatives to the next, or from one family (location) to the next. It could be that pairs of alleles at a number of loci, say, AAbbccDDee, subject to a sequence of environmental factors, say, FghiJ, are associated, all other things being equal, with the same outcome for the trait as are alleles aabbCCDDEE subject to a sequence of environmental factors FgHiJ (Taylor 2008). If underlying factors can be heterogeneous, the use of heritability as a

basis for judging a trait to be a good candidate for molecular research (Nuffield Council on Bioethics 2002, chapter 11) becomes unreliable (Taylor 2010). (Similarly for research that builds on the other fractions of the variance.) Underlying heterogeneity would help explain the difficulties genomic studies have had in identifying causally relevant genetic variants behind variation in human traits (McCarthy et al. 2008, Couzin-Frankel 2010)…

References

Couzin-Frankel, J. (2010). "Major Heart Disease Genes Prove Elusive." Science **328**(5983): 1220-1221.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, J. Hirschhorn (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nature Reviews Genetics, **9**, 356-369.

Nuffield Council on Bioethics (2002). "Genetics and Human Behavior: The Ethical Context." http://www.nuffieldbioethics.org, viewed 22 June 2007.

Taylor, P. J. (2008). "The under-recognized implications of heterogeneity: Opportunities for fresh views on scientific, philosophical, and social debates about heritability," History and Philosophy of the Life Sciences, **30**: 431-456.

Taylor, P. J. (2010). "Three puzzles and eight gaps: What heritability studies and critical commentaries have not paid enough attention to." Biology & Philosophy, **25**:1-31.

Taylor, P. J. (2011). "The results and interpretation of classical quantitative genetics under alternatives to three standard assumptions." Ms.