Chances are that the Odds are Wrong

The Sample of One

The recent shuttle disaster has produced a spate of statistical interpretations in the news media that once again demonstrate the common tendency to read more into statistics than may be warranted by the facts. That tendency is no stranger to orthodontics.

Wide-ranging estimates of the safety of the shuttle system, usually presented in terms indicating the likelihood of an accident, have ranged from one in a few hundred to one in thousands. Each "prediction" was probably based on actual tests of samples of critical parts, using accepted engineering and statistical methods to produce a final number.

Prediction in such uncharted waters is far from an exact science, and interpretation of the resulting predictions is even less exact. A common error in interpretation of probability (risk) estimates is exemplified by news commentaries suggesting that the tragic failure after 25 missions proves earlier more "optimistic" estimates to be wrong.

Wrong!

It would take a thousand missions to prove a 1:1,000 probability right or wrong, and even those thousand observations would be subject to further statistical evaluation for an estimate of what might be expected in the next thousand.

The misconception lies in the assumption that a 1:1,000 failure ratio means that safety is assured for the first 999 events, when in actual fact the single failure in that ratio would be just as likely to occur in the first event as in the 999th.

There is a parallel in such statistical values as $p<.05$, which tend to lull us into the assumption that the implied one case in twenty could not possibly be the one under scrutiny. It *can* be the one, and this value appears with sufficient frequency in many articles to make it very likely that they include at least one such exception where the "significant" is actually meaningless.

Sample Size

Many statistical values are extrapolations that go well beyond the sample. In fact, that is often the very reason for using statistical methods — to present a reasonably realistic broad picture when only a relatively small sample is available. The value $p<.001$ is rarely based on a thousand cases.

Large samples *do* smooth out the data spectrum, especially when there is only one randomly-distributed variable, like heads/tails. Various criteria have been established for determining an acceptable minimum sample size.

While such criteria are reasonable for certain applications, they are nevertheless arbitrary. One does not drop off a cliff at some minimum sample size; the level of "reliability" of the conclusions merely slips below some arbitrary minimum.

The quest for an "adequate" sample size can involve an important tradeoff that is usually ignored because the negative effects lie outside the variables being studied at the moment.

The Negative Effect of Large Sample Size

While large sample sizes smooth out the distribution curves for random events and some other values, they can also *introduce new variables* and *obscure individual variation.*

New Variables are a price often paid for large samples. Not enough subjects are available from one sex group, so sexes are combined. The age range may be expanded. Efforts to maintain discrete anatomic groups, such as malocclusion or small limits for certain cephalometric measurements, may have to be abandoned. Treatment factors such as appliance and operator may be broadened. Any such dilution of efforts to maintain constants can have unpredictable and usually unknown counterproductive effects on results.

Individual variation is the stuff that orthodontics is made of. We modify some variables to bring them into harmony with others that may not be so modifiable. Knowing which factors can be safely modified, and which cannot, is a very important part of orthodontic diagnosis that cannot be obtained from mass data.

Some of the variables that can be critical for the individual patient who puts their future in our hands may be swamped in the mass analysis of large samples. This effect is often aggravated when sample selection is based on one or two criteria, while all of the others range widely. We *do* need to know central trends and tendencies, but it is equally important to know and understand the deviations and potential deviations that do not fit a statistical mold.

Orthodontics is a one-on-one service, but one-on-one studies are often frowned on as "anecdotal."

The Sample of One

Going back to the shuttle, it was a sample of one that changed the whole perception, and the very future, of space travel. What if that cold front had not sagged quite so far south, and the flight had gone well? How long would it have been before a failure? What would have been accomplished in the interim? Would the effect of a failure after 100, or 1000, flights have been the same?

Or, what if that single failure had occurred on the first flight? We cannot begin to guess, or "predict," the overall effects on present and future space exploration.

The atomic bomb is another example. It took only one test to confirm the theoretical basis sufficiently to convince the decision-makers to move on to "practical" application. It did not require a hundred White Sands, or a hundred Hiroshimas, to arrive at some very firm conclusions.

Alternate Approaches

Observational data holds information that may be evaluated in many ways, of which mass statistical evaluations are only one type. Such statistics help us to see what is already there, but they do not produce new facts.

Mere lack of a "valid" statistical sample is is not a valid reason for failure to look at the wealth of information that is available for the looking.

With or without the availability of broad statistical evaluations to provide perspective, there is invariably a wealth of information in every single datum point — even in the sample of one. We orthodontists spend our lives treating samples of one, and we soon learn that the "average" patient is a myth. Even the individual subjects whose measurements have provided our "norms" did not fit those norms.

The problems of the patient confronting us at any given moment rarely match the central tendencies, and that patient's overriding concerns are what will happen to *them,* not some mythical average person who would not even require therapy. Means and standard deviations will tell us little about the condition of that patient's critical booster seals.

As responsible professionals, we cannot allow the scientific trappings of seemingly precise numbers to lure us away from looking with a critical eye at every unique detail and combination in that one-of-a-kind sample of one who puts their future into our hands.

Raymond C. Thurow