In this department *The American Statistician* publishes articles, reviews, and notes of interest to teachers of the first mathematical statistics course and of applied statistics courses. The department includes the Accent on Teaching Materials section; suitable contents for the section are described under the section heading. Articles and notes for the department, but not intended specifically for the section, should be useful to a substantial number of teachers of the indicated types of courses or should have the potential for fundamentally affecting the way in which a course is taught.

What About the Other Intervals?

STEPHEN B. VARDEMAN*

For a variety of introductory audiences, there are strong practical and pedagogical reasons for the early teaching of statistical interval methods that are often treated as "advanced" topics, if at all. There are also simple, effective ways of making this early introduction. This expository article discusses the elementary teaching of one-sided statistical intervals, prediction intervals, and tolerance intervals for both (one-sample) nonparametric and (general) normal theory contexts.

KEY WORDS: Confidence intervals; Nonparametric intervals; Normal theory intervals; One-sided intervals; Prediction intervals; Tolerance intervals.

1. INTRODUCTION

Standard introductions to data-based interval making (for audiences ranging from the typical Stat 101 "statistics for all comers" courses to both theory and methods courses for graduate statistics majors) consist primarily of discussions of *two-sided* normal theory *confidence intervals* for one- and two-sample situations. In those cases where regression is included in an introductory course, it sometimes happens that *prediction intervals* are taught in parallel with confidence intervals for a mean response, but curiously enough, typically only for the regression context. [Whitmore (1986) and Scheuer (1990) have both recently argued for fuller coverage of prediction intervals in our teaching.] A few engineering-oriented introductory texts mention *tolerance intervals*, but only in an almost obligatory fashion, and the topic seems completely lacking in more general introductions.

I believe that such treatments are lacking in at least two important respects. That is:

1. *They fail to address practical needs for applicable statistical tools* of even the most introductory "methods" audiences.

2. *They are pedagogically deficient* in that they fail to capitalize on important opportunities to reinforce other course material and to exposit some general principles of statistical practice, in that they leave students with some unfortunate misconceptions and confusion, and in that a number of matters of exposition could be smoothed out if a more complete introduction to statistical intervals were presented.

The purpose of this expository article is to elaborate slightly on these failings and to discuss a number of suggestions for improvement.

It is the thesis here that it is both possible and advisable to bring into the mainstream of statistical teaching topics such as: (a) one-sided intervals, (b) distribution-free intervals, (c) prediction intervals for all kinds of contexts, and (d) tolerance intervals for all kinds of contexts. These are presently treated as "advanced" or "secondary" topics, but they deserve to be taught *early* and as the important tools that they are. I have strong opinions about what material in current introductory courses ought to be dropped in order to make room for these topics, but will not address that issue here.

We will begin with a section reviewing the kinds of probability guarantees associated with the various types of statistical intervals and briefly discussing the different kinds of problems they address. We will then in turn comment on the teaching of one-sided intervals, distribution-free intervals, prediction intervals, and tolerance intervals.

2. TYPES OF STATISTICAL INTERVALS

For purposes of laying out the probability guarantees associated with various types of statistical intervals, suppose that one is furnished with n independent univariate continuous measurements $y_1, y_2, \ldots, y_n$. If, for example, one further supposes that there is a single common probability distribution F belonging to some class of distributions $\mathcal{F}$ generating these observations, and some parameter or functional of F, say Θ_F, is of interest, methods for using $y_1, y_2, \ldots, y_n$ to make *confidence intervals* for Θ_F are well known for a huge variety of versions of this problem. If we write $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and $I(\mathbf{y})$ is a random interval constructed from the

*Stephen B. Vardeman is Professor, Statistics and Industrial and Manufacturing Systems Engineering Departments, Iowa State University, Ames, IA 50011. The author is grateful for interaction with Gerry Hahn and Bill Meeker, who first prodded him in the direction of many of the views expressed in this article. He also wishes to thank the referee and associate editor for helpful reviews of an earlier draft of the article.

data, then it is completely standard to call $I(\mathbf{y})$ a $\gamma \times 100\%$ confidence interval for Θ_F provided

$$P_F[\Theta_F \in I(\mathbf{y})] \geq \gamma \quad (1)$$

for all $F \in \mathcal{F}$.

The "problem" with teaching only confidence intervals is that many practical problems are more appropriately phrased in terms of *individual measurements* rather than parameters of distributions. For most purposes, for example, a consumer should be far more interested in an EPA gas mileage test result applicable to a particular vehicle he or she is considering buying than the average figure associated with all such vehicles. Prediction and tolerance-interval methods address problems of inference for (future) individual measurements.

Continuing, for example, to consider the one-sample model where one's data $y_1, y_2, \ldots, y_n$ are iid F, if y_{n+1} is a single additional undisclosed observation independent of $\mathbf{y}$ and also having distribution F, $I(\mathbf{y})$ is commonly called a $\gamma \times 100\%$ prediction interval for y_{n+1} provided

$$P_F[y_{n+1} \in I(\mathbf{y})] \geq \gamma \quad (2)$$

for all $F \in \mathcal{F}$. It is important to recognize the nature of the probability calculation indicated in (2). The probability averaging is over the entire joint distribution of $(\mathbf{y}, y_{n+1})$. When prediction intervals are taught, unless extreme care is taken in exposition, students can be left with the misconception that the F probability content of $I(\mathbf{y})$ is guaranteed to be at least γ [i.e., that "a fraction γ of future observations will fall in $I(\mathbf{y})$"].

The F probability content of $I(\mathbf{y})$ is *random* and all that is really guaranteed by (2) in this direction is that, for example, with $I(\mathbf{y}) = [L(\mathbf{y}), U(\mathbf{y})]$,

$$E_F\{F[U(y)] - F[L(y)]\} \geq \gamma \quad (3)$$

for all $F \in \mathcal{F}$. The *expected* F probability content of $I(\mathbf{y})$ is at least γ, but the distribution of the F probability content can place substantial mass to the left of γ! (Many samples will lead to intervals covering less than $\gamma \times 100\%$ of the underlying distribution.)

The usual $\gamma \times 100\%$ prediction confidence guarantee of type (2) relates to repetitions of the whole process of generating $\mathbf{y}$, constructing $I(\mathbf{y})$, and generating a *single* additional observation y_{n+1}. There are other less well-known methods for predicting "at least k out of m" future observations (that might, for example, be useful to a small business owner considering the purchase of a few vehicles where mileage figures are of concern). For an entrance into this more specialized literature the reader is referred to the excellent review article of Hahn (1970) and the book of Hahn and Meeker (1991).

Tolerance-interval methods are meant to locate the bulk of an underlying distribution of individual measurements. Continuing to use the one-sample scenario as a basis for illustration, if $I(\mathbf{y}) = [L(\mathbf{y}), U(\mathbf{y})]$ the random interval $I(\mathbf{y})$ can be called a $\gamma \times 100\%$ tolerance interval for a fraction p of the underlying distribution (equivalently, all future observations) provided

$$P_F\{F[U(\mathbf{y})] - F[L(\mathbf{y})] \geq p\} \geq \gamma \quad (4)$$

for all $F \in \mathcal{F}$.

Comparing (3) and (4) we see that a 95% tolerance interval for a fraction .90 of an underlying distribution does with probability at least .95 what a 90% prediction interval does only on average. (It incidentally also does for all future observations what a 95% prediction interval for $k = 9$ out of $m = 10$ does for 10 specific values.) It should therefore be intuitively plausible that tolerance intervals for large fractions p and having usual confidence levels γ will typically be larger than corresponding $p \times 100\%$ prediction intervals. It is simply a more ambitious proposition to use data to bracket most of all future observations than to bracket a single future observation.

It may be helpful before leaving this section to think about the different uses to which confidence intervals, prediction intervals, and tolerance intervals produced from a single set of data might be put. So consider once again a proverbial EPA mileage test scenario, in which several nominally identical autos of a particular model are tested to produce mileage figures $y_1, y_2, \ldots, y_n$. If such data are processed to produce a 95% confidence interval for the mean mileage of the model, it is, for example, possible to use it to project the mean or total gasoline consumption for the manufactured fleet of such autos over their first 5,000 miles of use. Such an interval, would however, not be of much help to a person renting one of these cars and wondering whether the (full) 10-gallon tank of gas will suffice to carry him the 350 miles to his destination. For that job, a prediction interval would be much more useful. (Consider the differing implications of being "95% sure" that $\mu \geq 35$ as opposed to being "95% sure" that $y_{n+1} \geq 35$.) But neither a confidence interval for μ nor a prediction interval for a single additional mileage is exactly what is needed by a design engineer charged with determining how large a gas tank the model really needs to guarantee that 99% of the autos produced will have a 400-mile cruising range. What the engineer really needs is a tolerance interval for a fraction $p = .99$ of mileages of such autos.

3. ONE-SIDED INTERVALS

Statistical interval making is almost always introduced as a two-sided matter. Indeed, many elementary texts do not treat one-sided interval making. This is unfortunate for two reasons, one practical and the other pedagogical. In the first place, many practical statistical problems are inherently directional/one-sided. A production engineer is typically worried about inadequate material strength, not about material strength that is off-specification on the high side. His needs are most sensibly served by data-based lower bounds, not two-sided intervals.

On the pedagogy side, the fact that one-sided intervals are usually treated as an afterthought puts an unnecessary kink in elementary introductions of confidence intervals. That is, most academics have had unpleasant experiences trying to explain to a sea of blank faces first being introduced to confidence intervals, why in making a 95% (two-sided) confidence interval for μ, one makes use of a tabled 97.5 percentile. That matter is not central to the mission of introducing confidence intervals and can be easily avoided by beginning with one-sided intervals.

There are one-sided versions of most interval methods, and they should be taught early, not ignored or treated as afterthoughts.

4. DISTRIBUTION-FREE ONE-SAMPLE INTERVALS

Table 1 contains the weights of $n = 100$ newly minted U.S. pennies (measured to the nearest 10^{-4} g but reported only to the nearest .02 g) taken from W. J. Youden's NBS Special Publication 672 entitled *Experimentation and Measurement.*

It seems obvious that the three most natural statistical intervals associated with the single sample in Table 1 are $(2.99, \infty)$, $(-\infty, 3.21)$, and $(2.99, 3.21)$. Why then do we typically open our discussions of statistical intervals with something like

$$3.108 \pm 1.645 \frac{.043}{\sqrt{100}},$$

the large sample 90% two-sided confidence limits for μ? It would seem far more natural to instead discuss the meaning and use of the "distribution-free" one-sample interval formulas

$$(\min y_i, \infty), \qquad (5)$$

$$(-\infty, \max y_i), \qquad (6)$$

and

$$(\min y_i, \max y_i), \qquad (7)$$

which lead to the numerical intervals above.

The intervals (5) through (7) have a number of pedagogical virtues. Consider first their use as confidence intervals, for Θ_F the median of a continuous distribution F. ($\mathcal{F}$ in (1) is the class of continuous distributions.)

It is a useful class exercise in the binomial distribution to argue that the confidence level associated with (5) or (6) as an interval for a population median is

Table 1. Weights of 100 Newly Minted Pennies

Penny weight	Frequency
2.99	1
3.01	4
3.03	4
3.05	4
3.07	7
3.09	17
3.11	24
3.13	17
3.15	13
3.17	6
3.19	2
3.21	1

$$(1 - (.05)^n) \times 100\%. \qquad (8)$$

Similarly, the elementary probability/set theoretic notions of complementation and disjointness, the "addition rule," and the use of the binomial distribution can all be reinforced during a class discussion of the fact that the confidence level associated with (7) is

$$(1 - 2(.5)^n) \times 100\%. \qquad (9)$$

In the progress of the arguments leading to (8) and (9) it becomes obvious what "goes wrong" when the corresponding interval fails. Students are therefore also given a fairly concrete path to understanding the meaning of confidence.

The explicit Formulas (8) and (9) offer additional pedagogical advantages as well. For one thing, they enable explicit sample size calculations that do not depend upon "guessing" or bounding the values of any distribution parameters. As such, they provide a clean introduction to questions of the type, "How much data do I need to . . . ?" A related point is that (9) provides a reliability figure to attach to the skeletal box plots that we routinely teach early in elementary courses. [In fact, (9) combined with the notion of independence leads to a reliability figure to attach to a display of several side-by-side skeletal box plots.]

Besides thinking of Intervals (5), (6), and (7) as confidence intervals for a median, one can also consider their use as one-sample prediction and tolerance intervals. The most elementary probability arguments possible, involving sample spaces with equally likely outcomes (that incidentally have little else of practical use to recommend them) show that the prediction confidence to be associated with Intervals (5) and (6) is

$$\left(\frac{n}{n+1}\right) \times 100\%, \qquad (10)$$

while that associated with (7) is

$$\left(\frac{n-1}{n+1}\right) \times 100\%. \qquad (11)$$

The simple arguments leading to (10) and (11) again not only reinforce elementary probability concepts, but also make clear how the corresponding intervals "fail" and in the process show students what a prediction guarantee of the sort (2) means.

Treating Intervals (5) and (6) as tolerance intervals for a fraction p of an underlying distribution, an elementary binomial distribution argument again shows the associated confidence level to be

$$(1 - p^n) \times 100\%. \qquad (12)$$

A class discussion of the origin of (12) makes the similarity to (8) apparent and allows one to point out that one-sided tolerance intervals are in fact one-sided confidence intervals for the pth quantile [or $(1 - p)$th quantile] of F.

The confidence level associated with (7) as a tolerance interval for a fraction p of an underlying distribution is well known to be

$$(1 - p^n - n(1 - p)p^{n-1}) \times 100\%, \qquad (13)$$

and while the derivation of (13) is not elementary, the result is fairly explicit and does amount to a nice exercise in order statistics at the "introductory" graduate level.

Using Formulas (5), (6), and (7) in all of their roles as confidence, prediction, and tolerance intervals provides the opportunity to discuss the different purposes served by the various kinds of statistical intervals in a most concrete way. It also provides the opportunity to clearly contrast the different confidence levels (or equivalently the different sample size requirements for a fixed confidence) associated with the various applications. For example, (8), (10), and (12) applied to the situation of Table 1 show 2.99 g to be a lower confidence bound for the median weight of newly minted pennies with "unconfidence" on the order of 8×10^{-31}, to be a 99.01% lower prediction bound for a single additional penny weight, but to be only a 63% lower tolerance bound for 99.01% of all additional penny weights.

5. PREDICTION INTERVALS

We have already argued that some discussion of prediction intervals belongs in first statistics courses. We wish now to press the view that not only *some* but a *full* coverage of prediction interval methods is both possible and desirable.

One of the murkiest distinctions left in the minds of many graduates of elementary statistics courses is that between means and individuals. Anecdotal evidence compiled by this author in over 15 years of university teaching suggests that many of our students never really understand that

$$\bar{y} \pm t\frac{s}{\sqrt{n}}$$

is meant to bracket μ rather than y_{n+1}. Indeed they (consciously or not) seem to *want* a data-based interval for y_{n+1}. In quality control/engineering contexts this confusion manifests itself in an almost unbelievably persistent tendency to confuse process "control limits" (that have to do with the monitoring of μ via the use of $\bar{y}$) with "specifications" (that have to do with individual observations, y). The serious teaching of prediction intervals (and tolerance intervals too) seems one sensible way to try and combat this kind of confusion.

Spotty teaching of prediction intervals actually contributes its own special form of student misconceptions. That is, it seems that the only context in which most introductory students presently see prediction intervals is that of simple linear regression. This has the unfortunate effect of feeding students' tendency to expect "prediction" to have something to do with "extrapolation" or what one can expect to encounter under conditions completely unlike those that led to data in hand. A more complete teaching of prediction intervals is needed if this tendency is to be avoided.

The normal theory one-sample counterpart to Formulas (5), (6), and (7) is of course to use intervals based on one, the other, or both of the endpoints:

$$\bar{y} \pm ts\sqrt{1 + \frac{1}{n}} \quad . \qquad (14)$$

Comparing the use of (14) to that of a prediction interval application of (5), (6), or (7) for a "normal-looking" sample of moderate size is a clean way of introducing beginning students to the general principle that "the stronger the model assumptions under which one can operate, the sharper the resulting inferences."

A related point is that prediction (and tolerance) interval making is the natural context in which to expose students to the notion of using nonlinear transformations before applying normal theory methods. We routinely tell students to "transform, use standard formulas, and then untransform." This glib prescription has the unfortunate drawback that many of our favorite functionals (like means) do not survive nonlinear transformations. So for instance, transforming, making a normal theory confidence interval for a *mean*, and untransforming lead to the awkward interpretation of producing a confidence interval for a *median*. But no such awkwardness exists when one applies the prescription in the making of prediction (and tolerance) intervals. Untransforming prediction interval endpoints for a transformed measurement produces endpoints of a prediction interval for an additional raw value.

Normal theory prediction interval formulas generalizing (14) to two sample contexts, unstructured r sample contexts, simple linear regression, multiple linear regression, and factorial "few effects" contexts all follow quite simply from standard linear model theory. That is, it is clear that if x specifies conditions under which a prediction is desired (e.g., specifying levels of one or more factors) and under some model $\hat{y}_x$ is a normally distributed data-based predictor independent of an additional normal observation at conditions x, y_x, with $E(y_x - \hat{y}_x) = 0$, var $\hat{y}_x = A_x^2\sigma^2$ for a nonnegative constant A_x, var $y_x = \sigma^2$, and s^2 is an estimate of σ^2 independent of $\hat{y}_x$ and y_x possessing a $\sigma^2/\nu \cdot \chi^2_\nu$ distribution, endpoints of both one- and two-sided prediction intervals for y_x can be made using

$$\hat{y}_x \pm ts\sqrt{1 + A_x^2}, \qquad (15)$$

for t an appropriate quantile of the t distribution with ν degrees of freedom. In my opinion, for every linear model context that one is willing to teach students to do inference for mean responses, one ought also be willing to provide the corresponding version of (15). Symmetry/completeness alone would support this view even if the other arguments of this article were not considered.

6. TOLERANCE INTERVALS

We have already sprinkled through this discussion arguments to the effect that the early teaching of tolerance intervals is desirable. What really remains to say is what can be sanely done in "first" courses.

The normal theory tolerance interval counterpart to Formulas (5), (6), and (7) is to use intervals based on endpoints of the form

$$\bar{y} \pm \tau s. \tag{16}$$

Odeh and Owen (1980) present rather extensive tables of constants τ useful for both one- and two-sided tolerance intervals for various fractions p of an underlying distribution at various confidence levels $\gamma \times 100\%$. Intervals (16) should be a serious part of introductions to statistical methods.

But beyond (16) is the possibility of providing at least one-sided tolerance-interval methods for essentially *any* instance of the constant variance normal linear model. It is a first-year graduate level exercise to show that under the conditions leading to (15), a $\gamma \times 100\%$ one-sided tolerance bound for a fraction p of all future observations at conditions x can be made using one of the endpoints

$$\hat{y}_x \pm \tau_x s, \tag{17}$$

where

$$\tau_x = A_x \cdot Q_{t(\delta_x,\nu)}(\gamma) \tag{18}$$

for $Q_{t(\delta,\nu)}(\cdot)$ the inverse noncentral t cdf for noncentrality parameter δ and degrees of freedom ν, where $\delta_x = Q_z(p)/A_x$. [$Q_z(\cdot)$ is standing for the inverse standard normal cdf.]

These days many statistical packages can be used to provide the noncentral t quantile needed in (18) [for example, the SAS Supplemental Library function TINV(P, DF, NCT) could be used]. But there is also an old route to an explicit approximation of the limits (17). That is, it is a relatively simple exercise to show that under the conditions leading to (15), approximation of the distribution of $\hat{y}_x + ks$ by a normal distribution with mean $E\hat{y}_x + k\sigma$ and variance $\sigma^2(A_x^2 + k^2/2\nu)$ leads to the conclusion that approximate $\gamma \times 100\%$ one-sided tolerance bounds for a fraction p of all future observations at conditions x can be made using (17) and

$$\tau_x \approx \frac{Q_z(p) + A_x Q_z(\gamma)\sqrt{1 + \dfrac{1}{2\nu}\left(\dfrac{Q_z^2(p)}{A_x^2} - Q_z^2(\gamma)\right)}}{1 - \dfrac{Q_z^2(\gamma)}{2\nu}}. \tag{19}$$

This kind of approximation is common (at least in one-sample contexts) in the statistical quality control literature and is traceable to Jennett and Welch (1939).

Formula (19) is usually adequate for practical purposes. Thus, even if one does not have access to software needed to use (18), it is possible to specialize (17) to each one of the spectrum of linear models used in introductory courses, and to thus provide reasonably explicit normal theory tolerance bounds.

7. SUMMARY

There is, of course, no end to the list of statistical interval methods that one might potentially recommend for inclusion in a first course. (I even believe *simultaneous* interval methods to have a place!) Most readers will, however, probably conclude that the suggestions here go beyond what they are presently willing to attempt. But it is hoped that this discussion will provoke *some* thought and movement on the part of a number of instructors toward a more comprehensive early teaching of "the other intervals."

REFERENCES

Hahn, G. J. (1970), "Statistical Intervals for a Normal Population," *Journal of Quality Technology*, 2, 115–125; 195–206.

Hahn, G. J., and Meeker, W. Q. (1991), *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley.

Jennett, W. J., and Welch, B. L. (1939), "The Control of Proportion Defective as Judged by a Single Quality Characteristic Varying on a Continuous Scale," *Journal of the Royal Statistical Society*, Ser. B, 6, 80–88.

Odeh, R. E., and Owen, D. B. (1980), *Tables for Normal Tolerance Limits, Sampling Plans and Screening*, New York: Marcel Dekker.

Scheuer, E. M. (1990), "Let's Teach More About Prediction," in *Proceedings of the Statistical Education Section, American Statistical Association*, pp. 133–137.

Whitmore, G. A. (1986), "Prediction Limits for a Univariate Normal Observation," *The American Statistician*, 40, 141–143.

Jumping to Coincidences: Defying Odds in the Realm of the Preposterous

JAMES A. HANLEY*

1. INTRODUCTION

The calculation of probabilities is central to statistical inferences; however, teachers find it difficult to guide students, especially those who are not trained in probability, on how to set up correct probability calculations. Probabilities of seemingly rare events that are assessed *after the fact* are especially problematic; students employ selective vision and ignore other similar events in the sample space that would have prompted the same surprise and should therefore have been included in the calculated probability.

In a small publication aimed mainly at high school teachers, Hanley (1984) described three examples where probability specialists themselves have been "near-

*James A. Hanley is Associate Professor, Department of Epidemiology and Biostatistics, McGill University, Montreal PQ, H3A 1A2, Canada.

This work was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada. Josée Dupuis and Julie Bérubé provided helpful comments on the article.