datapoints
a statistics q&a

# Statistical Intervals

## Part 2: The Prediction Interval

BY STEPHEN N. LUKO AND DEAN V. NEUBAUER

**In this series of articles we continue to develop and review the statistical interval concept, here focusing on the prediction interval. As a means of demonstrating the idea of a prediction interval, we will utilize an example from the first article in this series to show a direct comparison.**

### Q: What is a prediction interval?

A: A prediction interval is an interval constructed using a set of sample data so as to contain a future observation(s). Note that this is a different problem than constructing an interval for the mean with some degree of confidence as shown in Part 1 of this article series. We assume that a future sample is taken under the same conditions and from the same population or process as the original sample and that the sample was random or the process was in a state of statistical control. There are many variations on this theme, but all are concerned with the essential problem of what will happen in the future and how often – the essence of statistics. We can have prediction intervals for variables data, or for attribute type data; we can further base the prediction on a parametric model such as the normal distribution or use nonparametric methods. Both are useful in practice. We can also place conditions on the future prediction. For example, we may want to have the interval contain at least 4 out of the next 5, or contain the mean from the next sample of 10. In this article we explore the common application of prediction intervals where the normal distribution applies.

Let us suppose that we have a random sample of $n$ observations $X_1, X_2, ..., X_n$ and we know that the data come from a normal distribution, but we do not know the mean and standard distribution of the distribution. A single future observation would be $X_{n+1}$ and its prediction error would be $X_{n+1} - \bar{x}$. The variance of this prediction error can be shown to be:

$$\sigma^2\left(1 + \frac{1}{n}\right) \text{ which is estimated by } s^2\left(1 + \frac{1}{n}\right) \qquad (1)$$

We want a prediction interval for the next single observation from this normal distribution. For our purposes, the formula is:

$$\bar{x} \pm t_{\alpha/2} s\sqrt{1 + \frac{1}{n}} \qquad (2)$$

The details of this theory may be found in Reference 1. The prediction interval for future observation $X_{n+1}$ will always be wider than a confidence interval for the mean μ due to the increased variability of the prediction error for a single observation versus the error of estimation for the mean. The term under the radical sign comes about because we are considering the variability in the sample average ($s/\sqrt{n}$) as well as the variability of the single future value ($s$). The value of $t_{\alpha/2}$ is a positive number taken from Student's $t$ distribution using $n$ - 1 degrees of freedom such that $P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha$. When actual numbers are substituted into Equation 2 we say that the resulting prediction interval has an associated confidence $C = 1 - \alpha$ of containing the next observation.

Recall the $n$ = 22 tensile adhesion tests made on U-700 alloy specimens. In Part 1, we found $\bar{x}$ = 13.71 and $s$ = 3.55 so the 95 percent confidence interval for μ was $12.14 \leq \mu \leq 15.28$. Applying the formula for this example using 95 percent confidence, the $t$ value using 21 degrees of freedom is $t$ = 2.080, and the prediction interval for the next observation, $X_{23}$, can be determined as follows.

$$\bar{x} - t_{\alpha/2,\, n-1} s\sqrt{1+\frac{1}{n}} \le X_{n+1} \le \bar{x} + t_{\alpha/2,\, n-1} s\sqrt{1+\frac{1}{n}}$$

$$13.74-(2.080)3.55\sqrt{1+\frac{1}{22}} \le X_{23} \le 13.74+(2.080)3.55\sqrt{1+\frac{1}{22}}$$

$$6.16 \le X_{23} \le 21.26$$

Notice the difference in the width of the prediction interval as compared to the confidence interval. Equation 2 is useful for situations where we may have a small data set, and data are scarce, such as for examples where we may get a value as seldom as one a week. In the case where the standard deviation, σ, is known, we substitute σ for *s* in Equation 2 and replace $t_{\alpha/2}$ by the standard normal quantile $Z_{\alpha/2}$. Suppose we want the interval to contain the next *k* observations. We only have to modify *t* in Equation 2. The interval for more than one future value must necessarily be larger than the interval for one future value because we are trying to capture multiple values at the same overall confidence level. There is an exact way to derive the modified *t* value, but most practitioners use the Bonferroni corrected *t* value. For a specified confidence $C = 1 - \alpha$, the *t* value is modified as $t_{\alpha/(2k)}$. For example, with 95 percent confidence and 21 degrees of freedom (in our example), an interval for the next five observations would use $t_{0.005}$ (i.e., $\alpha/(2k) = 0.05/(10) = 0.005$). This value is found to be 2.831. Using this value in Equation 2, the interval would be 3.43 to 23.99 and would then contain the next five observations with 95 percent confidence. We can do this for any number of future observations. In case we are interested in a one-sided prediction interval, the *t* value is adjusted as $t_{\alpha/k}$ (omitting the "2" in the subscript) for *k* the number of future values the interval is to contain. Now suppose we want a future one-sided interval for the next five observations at 95 percent confidence and that the interval is to be bounded on the high side. Here $\alpha = 0.05$, so use $t_{0.05/5} = t_{0.01}$ in Equation 2. For 21 degrees of freedom, $t_{0.01} = 2.518$. Since we want an upper bound we use the "+" form of Equation 2, giving 22.85 as the upper bound. Formally, the one-sided interval is (-∞, 22.85] with 95 percent confidence of containing the next five observations. There are many variations on this theme when the normal distribution applies.

It is important to note that the prediction interval is similar to a confidence interval in that the capture probability (confidence) is a long run result. That is, confidence is the long run proportion of cases, under the same conditions and with differing data, which would predict correctly what we say it would. For this and many other cases, including a comprehensive literature reference, readers are encouraged to review *Statistical Intervals: A Guide for Practitioners*, by Hahn and Meeker.[2]

## REFERENCES

1. Whitmore, G. A., "Prediction Limits for a Univariate Normal Observation," *The American Statistician*, May 1986, Vol. 40, No. 2.

2. Hahn, G. J., and Meeker, W. Q., *Statistical Intervals: A Guide for Practitioners*, Wiley-Interscience, John Wiley and Sons Inc., New York, N.Y., 1991.

**STEPHEN N. LUKO**, *Hamilton Sundstrand, Windsor Locks, Conn., is the immediate past chairman of Committee E11 on Quality and Statistics, and a fellow of ASTM International.*

**DEAN V. NEUBAUER**, *Corning Inc., Corning, N.Y., is an ASTM fellow; he serves as vice chairman of Committee E11 on Quality and Statistics, chairman of Subcommittee E11.30 on Statistical Quality Control, chairman of E11.90.03 on Publications and coordinator of the DataPoints column.*

*In the next article in this series, we will discuss tolerance intervals and their use.*

*Statistics play an important role in the ASTM International standards you write, and a panel of experts is ready to answer your questions about how to use statistical principles in ASTM standards. Please send your questions to* SN *Editor in Chief Maryann Gorman at mgorman@astm.org or ASTM International, 100 Barr Harbor Drive, P.O. Box C700, West Conshohocken, PA 19428-2959.*