Statistical Intervals: Nonparametric

Part 1

BY STEPHEN N. LUKO AND DEAN V. NEUBAUER

Q: How are nonparametric intervals applied and used?

A. In several previous articles of this column we have discussed confidence, prediction and tolerance intervals where the underlying distribution was of the normal type. In this article we develop these concepts further using nonparametric methods that do not assume an underlying normal distribution. We continue to assume that the sample is a random representation of a population or from a process in a state of statistical control.

We turn first to prediction-type intervals when we do not know the underlying distribution of the variable. In this scenario, the practitioner has a random sample of n observations taken from some population or process under study and would like to create an interval using the sample maximum and/or minimum that predicts one or more future values with some confidence C. Any such set of n observations will partition the support of the distribution into $n + 1$ compartments or blocks. For example, if we have four observations we get 4 + 1 = 5 compartments. This is depicted below in Figure 1, using an arbitrary distribution, for the case of $n = 4$. The subscripts in the diagram denote the ordered values in the sample: the so-called order statistics. Thus $x_{(1)}$ is the smallest value and $x_{(4)}$ the largest in this example, and so on.

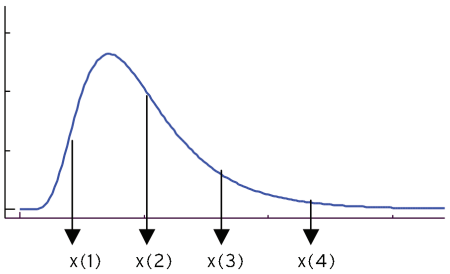

FIGURE 1 – Any distribution is partitioned into n + 1 compartments with a sample of size n. In this example $n = 4$.

Associated with each compartment is its probability (area under the curve) and, surprisingly, this may be shown to be $1/(n + 1)$ on average for all such compartments and for any distribution so partitioned. From these simple facts, we can estimate the probability that the next value will fall between any two order statistics. For example, if $n = 9$, the estimated probability that the next observation will fall below the largest observed value is approximated as $9/(9 + 1)$ or 0.9. It is common practice to call this estimated probability the "confidence" in the interval. This same confidence could also be applied to a future observation being greater than the smallest observed value in n. Both of these cases are examples of one-sided intervals. That is, the first interval is of the form $(-\infty, x_{(n)}]$ and the second of the form $[x_{(1)}, \infty)$, where $x_{(1)}$ and $x_{(n)}$ are the sample minimum and maximum values. One-sided intervals are very important in practice since many types of properties are bounded either as a maximum or a minimum.

If we wanted to use the same interval as a basis for predicting several, say k, future observations, then the formula for the associated confidence is adjusted as $C = n/(n + k)$ for either interval. For example, suppose $n = 22$, and we wanted to use the sample maximum as an upper bound for the next three observations and wanted to know the associated confidence in this interval. The answer is $C = 22/(22 + 3) = 0.88$, or 88 percent. It is not difficult to determine the sample size for a single-sided interval that would predict where the next k observations would fall. This is $n = kC/(1 - C)$. Thus, for 95 percent confidence and $k = 3$, we should use a sample size of 57. It is also a simple matter to create intervals using arbitrary order statistics, although intervals based on the sample minimum and maximum are more common.

Next, consider the case where the interval is constructed from the largest and smallest values

in the sample. This is an interval of the form [$x_{(1)}$, $x_{(n)}$]. Using this interval as a basis for predicting the next (k = 1) observation carries a confidence of $C = (n - 1)/(n + 1)$. Thus, if n = 22, then C = 21/23 = 0.913, or 91.3 percent. We summarize these basic cases below.

Case 1: The next k observations are less than the maximum (or the next k values are greater than the minimum) with confidence C.

$$C = \frac{n}{n+k} \qquad (1)$$

Case 2: The next single value falls between the sample minimum and maximum with confidence C.

$$C = \frac{n-1}{n+1} \qquad (2)$$

We can also calculate the confidence that the next k observations will fall between the maximum and the minimum of the original sample. The formula is not as intuitive as the single observation case, but by a subtle conditional probability argument it may be shown that the formula is Equation 3 below.

Case 3: The next k observations fall between the sample minimum and maximum.

$$C = \frac{n(n-1)}{(n+m)(n+m-1)} \qquad (3)$$

EXAMPLE

Suppose one has 39 observations and wants to use the sample minimum and maximum as a prediction interval for the next two observations. What is the confidence in using this interval? Using Equation 3, we find that C is 0.903, or approximately 90 percent. Note that in Equation 3, when m = 1, the formula reduces to Equation 2 for the single value prediction.

EXAMPLE

If we have 19 observations where the smallest value is 23.48 and the largest is 39.57, the confidence that the next observation lies between these two values is (19 - 1)/(19 + 1) = 0.9, or 90 percent confidence. The confidence that the next observation falls below the maximum (or above the minimum) is 19/(20 + 1) = 0.905, or 90.5 percent confidence.

EXAMPLE

Assuming we have n = 29 observations, what is the confidence that the next m = 3 observations fall between the minimum and the maximum of the original sample.

$$C = \frac{29(29-1)}{(29+3)(29+3-1)} = 0.819 \text{ or}$$

approximately 82 percent confidence.

We can also develop the confidence level for cases of prediction for k out of the next m observations, but these cases are less common and more complicated to use. Interested readers should consult *Mathematical Statistics* by S. S. Wilks for details.[1] 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 that would predict correctly what we say it would. For this and many other cases, including a comprehensive literature reference readers are encouraged to see *Statistical Intervals: A Guide for Practitioners*, by G. J. Hahn and W. Q. Meeker.[2]

REFERENCES

1. Wilks, S. S., *Mathematical Statistics*, John Wiley & Sons, New York, N.Y., 1963.
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, *United Technologies Aerospace Systems, Windsor Locks, Conn., is an ASTM fellow; a past chairman of Committee E11 on Quality and Statistics, he is current chairman of Subcommittee E11.30 on Statistical Quality Control.*

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

snonline

Get more tips for ASTM standards development at www.astm.org/sn-tips.

Find other DataPoints articles at www.astm.org/standardization-news/datapoints.