Statistical Intervals: Nonparametric

Part 2

BY STEPHEN N. LUKO AND DEAN V. NEUBAUER

Q: What considerations are there with nonparametric statistical intervals when the underlying distribution varies?

A. In Part 2 of this nonparametric statistical intervals series we consider tolerance type intervals where the underlying distribution can be of any type. We continue to assume that the sample is a random representative of a population or from a process in a state of statistical control.

NONPARAMETRIC TOLERANCE TYPE INTERVALS

A tolerance interval is an interval, one-sided or two-sided, constructed in a way to contain a specified proportion, p, of an entire population (distribution) with some confidence C. Tolerance intervals may apply to any kind of distribution, including the normal form. For details on the case of the normal distribution, see Reference 1.

Consider the case where 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 from a process under study and would like to create an interval using the sample maximum and/or minimum that predicts at least a proportion p of all future values, at some confidence level C. Suppose n is the sample size, and let $x_{(1)}$ and $x_{(n)}$ denote the sample minimum and maximum, respectively. There are three basic intervals of this type:

▶ Type 1, one-sided interval, Case 1: $[x_{(1)}, \infty)$,
▶ Type 1, one-sided interval, Case 2: $(-\infty, x_{(n)}]$, and
▶ Type 2, two-sided interval: $[x_{(1)}, x_{(n)}]$.

In each of these three cases we want to claim that the interval covers or contains at least a proportion p of the entire population that the data has come from, using a confidence level C. This is equivalent to stating that there is a confidence C that the probability is at least p that any future value of x would fall within the interval. In general, there is a relationship between n, p and C. Knowledge of any two of these three variables allows determination of the third.

For the one-sided cases $[x_{(1)}, \infty)$ or $(-\infty, x_{(n)}]$, we have essentially a success run of size n at or above the smallest order statistic (or at or below the largest order statistic). If we want to claim that at least a proportion p is greater than or equal to $x_{(1)}$, we have a success run of length n. This works like a binomial with probability p and n successes. The relationship is given as Equation 1a.

$$p \geq \sqrt[n]{1-c} \qquad \text{(1a)}$$

Equation 1a may be solved for p, n or C, which gives us two additional relationships.

$$C \geq 1 - p^n \qquad \text{(1b)}$$

$$n \geq \frac{ln(1-c)}{ln(1-p)} \qquad \text{(1c)}$$

For the question of sample size, use Equation 1c. Should we want to use 95 percent confidence and claim that a proportion of at least $p = 0.99$ lies above $x_{(1)}$, then using Equation 1c we find that $n = 299$ will just achieve this. Note that the two versions of the one-sided case are identical in this analysis. If we want to determine the confidence demonstrated for a specified

proportion and given sample size, use Equation 1b. For example, if we have $n = 22$ and want to claim that at least $p = 90$ percent of the population falls above the sample minimum, then $C \geq 1 - 0.9^{22} = 0.9015$ or approximately 90 percent confidence. Since the 10th percentile (also called the B10 life) is frequently required in materials and components testing, and since 90 percent confidence is in common use, one often sees $n = 22$ as a required sample size in materials or component testing.

For the two-sided case $[x_{(1)}, x_{(n)}]$, at least $100p$ percent of the population lies in the interval with confidence C, when a sample size of n is used. Analysis of this case leads to Equation 2 involving C, p and n. Details of the derivation are given in *Mathematical Statistics*, by S. S. Wilks.[2]

$$np^{n-1} - (n - 1) p^n \geq 1 - C \qquad (2)$$

This equation is solved iteratively for the unknown quantity when two of p, n and C are specified. Table 1 illustrates how this plays out among p, n and C. The table shows the sample size required at confidence level C to claim that the largest and smallest order statistics will bound a proportion of at least p of the population. For example, if we use $n = 130$, then with 99 percent confidence we can claim that at least 95 percent of future output will fall within the sample minimum and maximum.

We can also develop tolerance intervals using any arbitrary order statistics, but the most common interval uses the sample minimum and/or maximum values. Interested readers should consult the reference by S. S. Wilks[2] for details. It is important to note that tolerance intervals behave in much the same way as confidence and prediction type intervals. That is, the capture probability (confidence) is a long run result. In other words, 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 Hahn and Meeker.[3]

REFERENCES

1. Luko, Stephen and Neubauer, Dean., "Statistical Intervals Part 3: More on the Tolerance Interval" (DataPoints), ASTM *Standardization News*, November/December 2011.
2. Wilks, S. S., *Mathematical Statistics*, John Wiley & Sons, New York, N.Y., 1963.
3. 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.

Table 1 — Sample Size to Achieve a Confidence *C* that the Sample Minimum and Maximum Capture at Least *p* Percent of a Population or Process

	Confidence Level, %		
p%	90	95	99
99.9	3,889	4,742	6,635
99.0	388	473	662
98.0	194	236	330
97.0	129	157	219
96.0	96	117	164
95.0	77	93	130
94.0	64	78	108
93.0	55	66	92
92.0	48	58	81
91.0	42	51	71
90.0	38	46	64
85.0	25	29	41
80.0	18	21	30
75.0	15	17	23
70.0	12	14	19
65.0	9	11	16
60.0	8	10	13
55.0	7	8	11
50.0	6	7	10

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.*

EXAMPLE 1

If we have 37 observations from a materials breaking strength application and the minimum value in the sample is 1200, then we can claim with 95 percent confidence that a proportion of at least approximately 92 percent of the population lies at or above the sample minimum. Here, we have used Equation 1a with $n = 37$ and $C = 0.95$.

EXAMPLE 2

What sample size should we use if we want to be 90 percent confident that the sample minimum and maximum bounds at least 99 percent of a population? Use Equation 2 with $C = 0.9$ and $p = 0.99$ and increment n until the requirement of Equation 2 is just met. We find that $n = 388$ just meets the requirement.

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.