Statistical Prediction Intervals

Applying the Intervals to Attribute Data

BY STEPHEN N. LUKO AND DEAN V. NEUBAUER

Q: What considerations are there with statistical prediction intervals when the underlying distribution is known to be based on attribute data?

A. In our two previous DataPoints articles, we discussed nonparametric statistical intervals that are not dependent on an underlying statistical distribution.[1,2] This article concerns prediction intervals for the next observation when we have a set of data and our data is of the attribute type. There are two common cases. Case 1 occurs where the type of data we have is governed by the binomial distribution, and Case 2 occurs where the data are governed by the Poisson distribution. The intervals presented are approximate and are based on an approximating normal distribution. They should be useful for most cases where the initial sample observation is at least five events for the binomial case, and 10-15 or more for the Poisson case. The entire theory has been summarized previously by Hahn, Meeker and Nelson (see References 3 and 4). As in previously discussed interval estimation procedures, we continue to assume that all samples are a random representative of a population or from a process in a state of statistical control.

PREDICTION INTERVALS FOR THE BINOMIAL DISTRIBUTION

For attribute type data, the binomial distribution is one of the most important and widely applied in all of statistical practice. It is used where there is a fixed "event" probability p, a sample size n and a random variable r equal to the number of items in the sample that have the defined characteristic for the "event." The probability p is called a "success" probability but need not be a desirable type event. In the prediction interval context, the value of p is unknown. For n objects in a sample, one can observe at least 0 and at most n "successes." Often a "success" event is related to a quality attribute such as not meeting a requirement. Practitioners also call this a go/no-go type of sampling.

The problem may be stated as follows. We have an initial sample of size n and have observed r "events" among n inspections. In a future sample of size m, we will observe some number of events y. It is desirable to construct an interval that would contain y with some stated confidence, say C. The interval is called a prediction interval for the future observation y. Let $\hat{p} = r/n$ be the estimate of the unknown process average p, based on the initial sample size n. Let m be the future sample size, and let the confidence coefficient be $C = 1 - \alpha$. The following formula is used to construct the two-sided prediction interval for future number of events y.

$$m\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{m\hat{p}(1-\hat{p})(m+n)}{n}} \qquad (1)$$

In Equation 1, the quantity $Z_{\alpha/2}$ is a quantile selected from the standard normal distribution that leaves an area of $\alpha/2$ to the right of $Z_{\alpha/2}$. Thus, if 95 percent confidence is desired, $\alpha = 0.05$ and $Z_{0.025} = 1.96$. Equation 1 is derived from the fact that the estimate $\hat{p}$ will have a normal distribution in repeated application, as long as the number of observed events in the initial sample is five or more. For further details, see Reference 3 or 4.

EXAMPLE 1

A quality metric for a certain operation in a large firm is to measure the number of rejected material lots received by the firm's receiving inspection operation. This information is measured and reported to management on a monthly basis. The recent record indicates that last month seven of 107 shipments were rejected. Next month, the firm expects to receive 84 lots. As-

This article concerns prediction intervals for the next observation when we have a set of data and our data is of the attribute type. There are two common cases. Case 1 occurs where the type of data we have is governed by the binomial distribution, and Case 2 occurs where the data are governed by the Poisson distribution.

suming that the quality of incoming lots remains the same, what number y of rejected lots do we predict to occur in next month's inspections with 90 percent confidence? Here, n = 107, m = 84, $\hat{p}$ = 7/107 = 0.0654, α = 0.1 and $Z_{0.05}$ = 1.645. Using Equation 1, the resulting prediction interval for y is:

$$84(0.0654) \pm 1.645\sqrt{\frac{84(0.0654)(1-0.0654)(84+107)}{107}}$$

5.42 ± 4.95 or about 0.47 to 10.37.

We round this result to whole numbers as $0 \le y \le 11$. Thus we may expect between 0 and 11 as long as the process remains in statistical control, and the unknown process average p does not change.

PREDICTION INTERVALS FOR THE POISSON DISTRIBUTION

For the Poisson distribution, observations are made on an inspection region that can be based on time, area, space, number of objects or some other region description. The number of events we observe can be any whole number at least 0. The unknown parameter in this distribution is the rate of event occurrence λ. If r events are observed in an initial region of size s, the estimate of the rate is $\lambda = r/s$. In a future inspection region of size t, we will observe some number y of events. It is desired to construct a prediction interval for the variable y.

In this method, we assume that r is at least 10-15 or more. This assures that the statistic will be approximately normally distributed. We continue to use a confidence coefficient of C = 1

(CONTINUED) - α. The prediction interval for y, the number of future events in the region of size t, is constructed according to Equation 2 below.

$$\hat{\lambda}t \pm Z_{\alpha/2}\sqrt{\frac{\hat{\lambda}t(s+t)}{s}} \qquad (2)$$

EXAMPLE 2

In planning for future replacements of a certain component used on cell phone towers, a company would like to use the past two years of data in making a prediction for next year. In the last 24 months there were 29 replacements required. If things remain the same, how many replacements may we expect in the next year with 95 percent confidence?

Here, the current data comes from an interval of length s = 24 months. The observed number of events in this period is r = 29. The rate estimate is therefore λ = 29/24 = 1.208 events per month. The future period is t = 12 months and the confidence desired is C = 95 percent. The value of Z is $Z_{0.025}$ = 1.96. Using Equation 2, the prediction interval is constructed as:

$$1.208(12) \pm 1.96\sqrt{\frac{1.208(12)(24+12)}{24}}$$

14.5 ± 4.7 or about 9.8 to 19.2.

We round this result to whole numbers as $10 \le y \le 19$. Thus we may expect approximately between 10 and 19 replacements in the next 12 months so long as the process remains in statistical control and the process average (rate) does not change. It is important to note that in working with rates, s, t and the rate estimate have to have the same units in order to use Equation 2. In this example, the units were months.

Readers interested in the authors' three-part series on statistical intervals based on the normal distribution should see the DataPoints columns on statistical intervals in ASTM *Standardization News*, July//Aug. 2011, Sept./Oct. 2011 and Nov./Dec. 2011.[5-7]

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.*

REFERENCES

1. Luko, S.N. and Neubauer, D.V., "Statistical Intervals: Nonparametric, Part 1," ASTM *Standardization News*, Vol. 41, No. 6, Nov./Dec. 2013, pp. 20-21.
2. Luko, S.N. and Neubauer, D.V., "Statistical Intervals: Nonparametric, Part 2," ASTM *Standardization News*, Vol. 42, No. 1, Jan./Feb. 2014, pp. 12-13.
3. Hahn, Gerald J. and Meeker, William Q., *Statistical Intervals, A Guide for Practitioners*, Wiley InterScience, John Wiley and Sons Inc., New York, N.Y., 1991.
4. Hahn, G.J. and Nelson, Wayne, "A Survey of Prediction Intervals and Their Applications," Journal of Quality Technology, Vol. 5, No. 4, Oct. 1973.
5. Luko, S.N. and Neubauer, D.V., "Statistical Intervals, Part 1," ASTM *Standardization News*, Vol. 39, No. 4, July/Aug. 2011, pp. 18-20.
6. Luko, S.N. and Neubauer, D.V., "Statistical Intervals, Part 2," ASTM *Standardization News*, Vol. 39, No. 5, Sept./Oct. 2011, pp. 14-15.
7. Luko, S.N. and Neubauer, D.V., "Statistical Intervals, Part 3," ASTM *Standardization News*, Vol. 39, No. 6, Nov./Dec. 2011, pp. 18-19.

CORRECTION: In the January/February 2014 DataPoints column, "Statistical Intervals: Nonparametric, Part 2," Equation 1c should have been:

$$n \ge \ln(1 - C)/\ln(p)$$

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.