USGS
science for a changing world

Techniques of Water-Resources Investigations of the United States Geological Survey

Book 4, Hydrologic Analysis and Interpretation

Chapter A3

Statistical Methods in Water Resources

By D.R. Helsel and R.M. Hirsch

3.6 Parametric Prediction Intervals

Parametric prediction intervals are also used to determine whether a new observation is likely to come from a different distribution than previously-collected data. However, an assumption is now made about the shape of that distribution. This assumption provides more information with which to construct the interval, as long as the assumption is valid. If the data do not approximately follow the assumed distribution, the prediction interval may be quite inaccurate.

3.6.1 Symmetric Prediction Interval

The most common assumption is that the data follow a normal distribution. Prediction intervals are then constructed to be symmetric around the sample mean, and wider than confidence intervals on the mean. The equation for this interval differs from that for a confidence interval around the mean by adding a term $\sqrt{s^2} = s$, the standard deviation of individual observations around their mean:

$$\boxed{\mathrm{PI} = \bar{X} - t_{(\alpha/2,\, n-1)} \bullet \sqrt{s^2 + (s^2/n)} \quad \text{to} \quad \bar{X} + t_{(\alpha/2,\, n-1)} \bullet \sqrt{s^2 + (s^2/n)} \qquad [3.12]}$$

One-sided intervals are computed as before, using α rather than $\alpha/2$ and comparing new data to only one end of the prediction interval.

Example 2, cont.

Assuming symmetry, is a concentration of 350 ppb different (not just larger) than what would be expected from the previous distribution of arsenic concentrations? Use $\alpha = 0.10$.

The parametric two-sided $\alpha = 0.10$ prediction interval is

$$98.4 - t_{(.05,\, 24)} \bullet \sqrt{144.7^2 + 144.7^2/25} \quad \text{to} \quad 98.4 + t_{(.05,\, 24)} \bullet \sqrt{144.7^2 + 144.7^2/25}$$

$$98.4 - 1.711 \bullet 147.6 \quad \text{to} \quad 98.4 + 1.711 \bullet 147.6$$

$$-154.1 \quad \text{to} \quad 350.9$$

350 ppb is at the upper limit of 350.9, so the concentration is not declared different at $\alpha = 0.10$. The negative concentration reported as the lower prediction bound is a clear indication that the underlying data are not symmetric, as concentrations are non-negative. To avoid an endpoint as unrealistic as this negative concentration, an asymmetric prediction interval should be used instead.

3.6.2 Asymmetric Prediction Intervals

Asymmetric intervals can be computed either using the nonparametric intervals of section 3.5, or by assuming symmetry of the logarithms and computing a parametric interval on the logs of the data. Either asymmetric interval is more valid than a symmetric interval when the underlying data are not symmetric, as is the case for the arsenic data of example 2. As stated in Chapter 1, most water resources data and indeed most environmental data show positive skewness. Thus they should be modelled using asymmetric intervals. Symmetric <u>prediction</u> intervals should be used only when the data are known to come from a normal distribution. This is because prediction intervals deal with the behavior of individual observations. Therefore the Central Limit Theorem (see first footnote in this chapter) does not apply. Data must be assumed non-normal unless shown otherwise. It is difficult to disprove normality using hypothesis tests (Chapter 4) due to the small sample sizes common to environmental data sets. It is also difficult to see non-normality with graphs unless the departures are strong (Chapter 10). It is unfortunate that though most water resources data sets are asymmetric and small, symmetric intervals are commonly used.

An asymmetric (but parametric) prediction interval can be computed using logarithms. This interval is parametric because percentiles are computed assuming that the data follow a lognormal distribution. Thus from equation 3.12:

$$\mathrm{PI} = \exp\left(\bar{y} - t_{(a/2,n-1)}\sqrt{s_y^2 + s_y^2/n}\right) \text{ to } \exp\left(\bar{y} + t_{(a/2,n-1)}\sqrt{s_y^2 + s_y^2/n}\right)$$

where y = ln(X), $\bar{y}$ is the mean and s_y^2 the variance of the logarithms. [3.13]

<u>Example 2, cont.</u>

An asymmetric prediction interval is computed using the logs of the arsenic data. A 90% prediction interval becomes

$$\ln(\mathrm{PI}):\ 3.17 - t_{(0.05,\,24)} \bullet \sqrt{1.96^2 + 1.96^2/25} \quad \text{to} \quad 3.17 + t_{(0.05,\,24)} \bullet \sqrt{1.96^2 + 1.96^2/25}$$

$$3.17 - 1.71 \bullet 2.11 \quad \text{to} \quad 3.17 + 1.71 \bullet 2.11$$

$$0.44 \text{ to } 6.78$$

which when exponentiated into original units becomes

$$1.55 \quad \text{to} \quad 880.1$$

As percentiles can be transformed directly from one measurement scale to another, the prediction interval in log units can be directly exponentiated to give the prediction interval in original units. This parametric prediction interval differs from the one based on sample percentiles in that a lognormal distribution is assumed. The parametric interval would be preferred if the assumption of a lognormal distribution is believed. The sample percentile interval would be preferred when a robust interval is desired, such as when a lognormal model is not believed, or when the scientist does not wish to assume any model for the data distribution.