Lecture 3 Prediction Intervals, Tolerance Intervals and Control Charts

Suppose that, in a soil contamination monitoring process, new observations are available. If the new values greatly exceed the background or standard value, is this evidence of a true difference (i.e., is there contamination)? Or are the true underlying concentrations the same as background or the standard value and this is just a "chance" event?

One way to establish an objective decision rule to decide whether there is contamination or not is to base it on some statistical intervals. This decision rule is a specific kind of hypothesis test. In this talk, we discuss three statistical tools used to create background intervals: ***prediction intervals***, ***tolerance intervals***, and ***control charts***. Thus, the decision rule about whether contamination has occurred is based on whether the new observations fall inside or outside the background interval. Note that, in the following numerical examples, the packages {EnvStats} and {qcc} are required.

3.1 Prediction Interval

A prediction interval for some population is an interval on the real line constructed so that it will contain k future observations from that population with some specified probability $(1-\alpha)100\%$ (confidence level), where α is some fraction between 0 and 1 (usually less than 0.5), and k is some positive integer.

The basic idea of a prediction interval is to assume a particular probability distribution, e.g., normal, for some process generating the data, e.g., observations of chemical concentrations in soil, compute sample statistics from a baseline sample, and then use these sample statistics to construct a prediction interval, assuming the distribution of the data does not change in the future.

Example 3.1 The table below shows arsenic concentrations (ppb) collected quarterly at two groundwater monitoring wells (data in "*Arsenic.csv*").

Well	Year	Observed Arsenic (ppb)
Background	1	12.6 30.8 52.0 28.1
	2	33.3 44.0 3.0 12.8
	3	58.1 12.6 17.6 25.3
Compliance	4	48.0 30.3 42.5 15.0
	5	47.6 3.8 2.6 51.9

We use the data from the background well to construct a prediction interval for the next k = 4 observations, assuming that arsenic concentration is normally distributed (a valid assumption?).

```
predIntNorm(Arsenic$Background, n.mean = 1, k = 4, method = "Bonferroni",
        pi.type = "upper", conf.level = 0.95)
```

```
Results of Distribution Parameter Estimation
--------------------------------------------

Assumed Distribution:            Normal

Estimated Parameter(s):          mean = 27.51667
                                 sd   = 17.10119

Estimation Method:               mvue

Data:                            Arsenic$Background

Sample Size:                     12

Prediction Interval Method:      Bonferroni

Prediction Interval Type:        upper

Confidence Level:                95%

Number of Future Observations:   4

Prediction Interval:             LPL =     -Inf
                                 UPL = 73.67237
```

Since all 4 observations from the compliance well in year 4 (or year 5) are below the UPL, we conclude that there is no arsenic contamination in the year. ∎

Considering that the measurement in Example 3.1 takes only non-negative values and is usually positively skewed in distribution, we may fit a ***lognormal distribution*** to the data.

A positive-valued random variable X is said to have a lognormal probability distribution, if the random variable log(X) has a normal probability distribution. Figure 3.1 shows a lognormal density function.

Example 3.2 Re Example 3.1, we use the data from the background well to construct a prediction interval for the next k = 4 observations, assuming that the distribution of arsenic concentration is lognormal.

```
predIntLnorm(Arsenic$Background, n.geomean = 1, k = 4, method = "Bonferroni",
        pi.type = "upper", conf.level = 0.95)
Results of Distribution Parameter Estimation
--------------------------------------------

Assumed Distribution:            Lognormal

Estimated Parameter(s):          meanlog = 3.0733829
                                 sdlog   = 0.8234277

Estimation Method:               mvue

Data:                            Arsenic$Background

Sample Size:                     12

Prediction Interval Method:      Bonferroni

Prediction Interval Type:        upper

Confidence Level:                95%

Number of Future Observations:   4

Prediction Interval:             LPL =   0.0000
                                 UPL = 199.4961
```

All 4 observations from the compliance well in year 4 (or year 5) are clearly below the UPL, we thus conclude that there is no sign of contamination in the year. ■

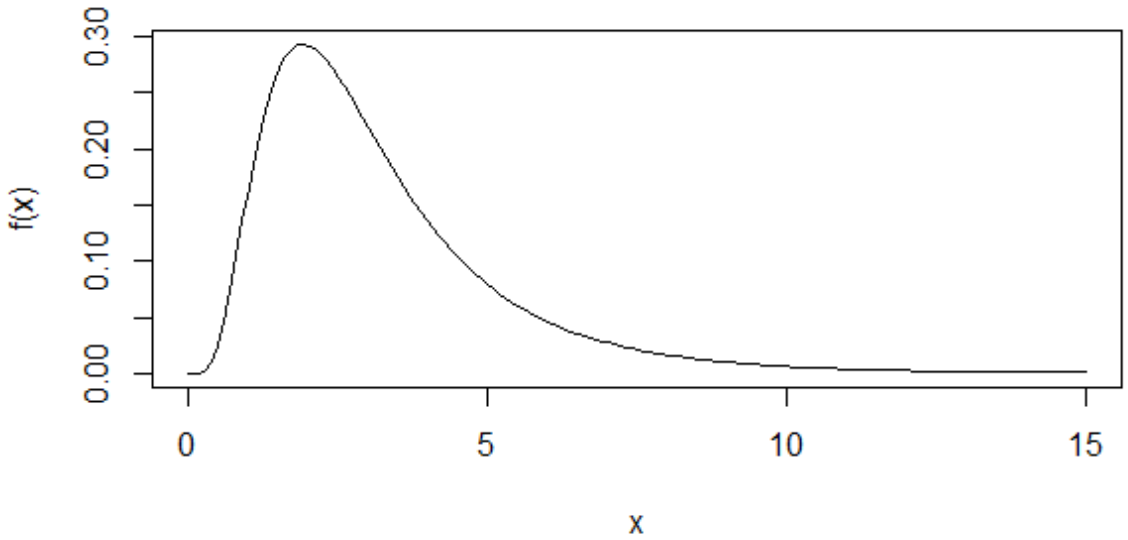


Figure 3.1 Lognormal probability density function.

3.2 Tolerance Interval

A ***tolerance interval*** for some population is an interval on the real line constructed so as to contain β100% of the population (i.e., β100% of all future observations), where $0 < \beta < 1$ (usually β is bigger than 0.5). The quantity β100% is called the ***coverage***. As with a prediction interval, the basic idea of a tolerance interval is to assume a particular probability distribution (e.g., normal, lognormal, etc.) for some process generating the data (e.g., quarterly observations of chemical concentrations in groundwater), compute sample statistics from a baseline sample, and then use these sample statistics to construct a tolerance interval, assuming the distribution of the data does not change in the future. In the case when the distribution of X is known, a β100% tolerance interval is exactly the same as a (1- α)100% prediction interval for k = 1 future observation, where $\beta = 1 - \alpha$.

There are two ways to construct tolerance intervals:

- A ***β-content*** tolerance interval with confidence level (1-α)100% is constructed so that it contains at least β100% of the population, i.e., the coverage is at least β100%, with probability (1- α)100%.
- A ***β-expectation*** tolerance interval is constructed so that it contains on average β100% of the population, i.e., the average coverage is β100%.

Prediction and tolerance intervals have long been applied to quality control and life testing problems. In environmental monitoring, tolerance intervals are used in two different ways,

- **Compliance-to-Background Comparisons**: Construct a tolerance interval based on background data, then compare data from a compliance well or site to the tolerance interval. If any compliance data are outside of the tolerance interval, then declare contamination is present.
- **Compliance-to-Fixed Standard Comparisons**: Construct a tolerance interval based on compliance data, then compare the tolerance limit to a fixed standard. If the tolerance limit is greater (less) than the fixed standard, declare contamination is present.

Example 3.3 Use upper tolerance limits (β-content and β-expectation) and an upper prediction limit to determine contamination at a clean-up site. The TcCB (1,2,3,4-

Tetrachlorobenzene) data are shown in the following table (data in *EPA.94b.tccb.df* of {EnvStats}).

Area	Observed TcCB (ppb)
Reference	0.22 0.23 0.26 0.27 0.28 0.28 0.29 0.33 0.34 0.35 0.38 0.39 0.39 0.42 0.42 0.43 0.45 0.46 0.48 0.50 0.50 0.51 0.52 0.54 0.56 0.56 0.57 0.57 0.60 0.62 0.63 0.67 0.69 0.72 0.74 0.76 0.79 0.81 0.82 0.84 0.89 1.11 1.13 1.14 1.14 1.20 1.33
Clean-up	ND 0.09 0.09 0.12 0.12 0.14 0.16 0.17 0.17 0.17 0.18 0.19 0.20 0.20 0.21 0.21 0.22 0.22 0.22 0.23 0.24 0.25 0.25 0.25 0.25 0.26 0.28 0.28 0.29 0.31 0.33 0.33 0.33 0.34 0.37 0.38 0.39 0.40 0.43 0.43 0.47 0.48 0.48 0.49 0.51 0.51 0.54 0.60 0.61 0.62 0.75 0.82 0.85 0.92 0.94 1.05 1.10 1.10 1.19 1.22 1.33 1.39 1.39 1.52 1.53 1.73 2.35 2.46 2.59 2.61 3.06 3.29 5.56 6.61 18.40 51.97 168.64

tolIntLnorm(TcCB[Area=="Reference"], coverage=0.95,cov.type="content", ti.type="upper", conf.level=0.95)

```
Results of Distribution Parameter Estimation
--------------------------------------------

Assumed Distribution:            Lognormal

Estimated Parameter(s):          meanlog = -0.6195712
                                 sdlog   =  0.4679530

Estimation Method:               mvue

Data:                            TcCB[Area == "Reference"]

Sample Size:                     47

Tolerance Interval Coverage:     95%

Coverage Type:                   content

Tolerance Interval Method:       Exact

Tolerance Interval Type:         upper

Confidence Level:                95%

Tolerance Interval:              LTL = 0.00000
                                 UTL = 1.42497
```

```
tolIntLnorm(TcCB[Area=="Reference"], coverage=0.95,cov.type="expectation", ti.type="upper")
Results of Distribution Parameter Estimation
--------------------------------------------

Assumed Distribution:            Lognormal

Estimated Parameter(s):          meanlog = -0.6195712
                                 sdlog   =  0.4679530

Estimation Method:               mvue

Data:                            TcCB[Area == "Reference"]

Sample Size:                     47

Tolerance Interval Coverage:     95%

Coverage Type:                   expectation

Tolerance Interval Method:       Exact

Tolerance Interval Type:         upper

Tolerance Interval:              LTL = 0.000000
                                 UTL = 1.190384
```

```
predIntLnorm (TcCB[Area=="Reference"], k=77, method="exact", pi.type="upper", conf.level=0.95)
Results of Distribution Parameter Estimation
--------------------------------------------

Assumed Distribution:            Lognormal

Estimated Parameter(s):          meanlog = -0.6195712
                                 sdlog   =  0.4679530

Estimation Method:               mvue

Data:                            TcCB[Area == "Reference"]

Sample Size:                     47

Prediction Interval Method:      exact

Prediction Interval Type:        upper

Confidence Level:                95%

Number of Future Observations:   77

Prediction Interval:             LPL = 0.000000
                                 UPL = 2.681076
```

Note that a lognormal distribution is assumed for the TcCB. Based on the limits we conclude that contamination is present at the clean-up site. ■

3.3 Control Charts

Control charts are a graphical and statistical method of assessing the performance of a system over time. They were developed in the 1920s by Walter Shewhart, and have been employed widely in industry to maintain process control (e.g., manufacturing a part for a car, airplane, or computer to within certain specifications). In the context of, for instance, groundwater monitoring, they have been suggested as an alternative to prediction or tolerance limits for monitoring constituent concentrations at compliance wells when enough historical data are available at each compliance well to establish reliable background values for each well.

Control charts assume the observations at a particular compliance well are independent and follow a normal distribution with some constant mean μ and standard deviation σ.

A ***Shewhart control chart*** is to plot the observations over time and compare them to established upper and/or lower control limits that are based on historical data. Once a single observation falls outside the control limit(s), this is an indication that the process is "out of control" and needs to be investigated.

Letting $\overline{x}$ and *s* denote the sample mean and standard deviation from the historical data, the upper and lower control limits then become

$$\text{UCL} = \overline{x} + L\text{s}$$
$$\text{LCL} = \overline{x} - L\text{s}.$$

The constant *L* is often set to *L* = 3, and then the *UCL* and *LCL* are called "3-sigma control limits".

To detect a gradual trend in the process, we may Cumulative Summation (CUSUM) charts. As its name suggests, a CUSUM chart involves cumulative sums. For the *i*th future sampling occasion, the *i*th upper cumulative sum S_i^+ and lower cumulative sum S_i^- by

$$S_i^+ = \max\left[0, \left(\frac{x_i - \overline{x}}{s} - k\right) + S_{i-1}^+\right], \text{ with } S_0^+ = 0,$$

$$S_i^- = \min\left[0, \left(-\frac{x_i - \overline{x}}{s} - k\right) + S_{i-1}^-\right], \text{ with } S_0^- = 0,$$

where *k* denotes a positive reference value that must be set by the user and corresponds to half the size of a linear trend (in units of standard deviations) deemed worthy of detecting quickly. Usually, we use *k* = 1, i.e., it is important to detect a trend of two standard deviations quickly. With a CUSUM chart, we declare a process "out of control" when the upper/lower

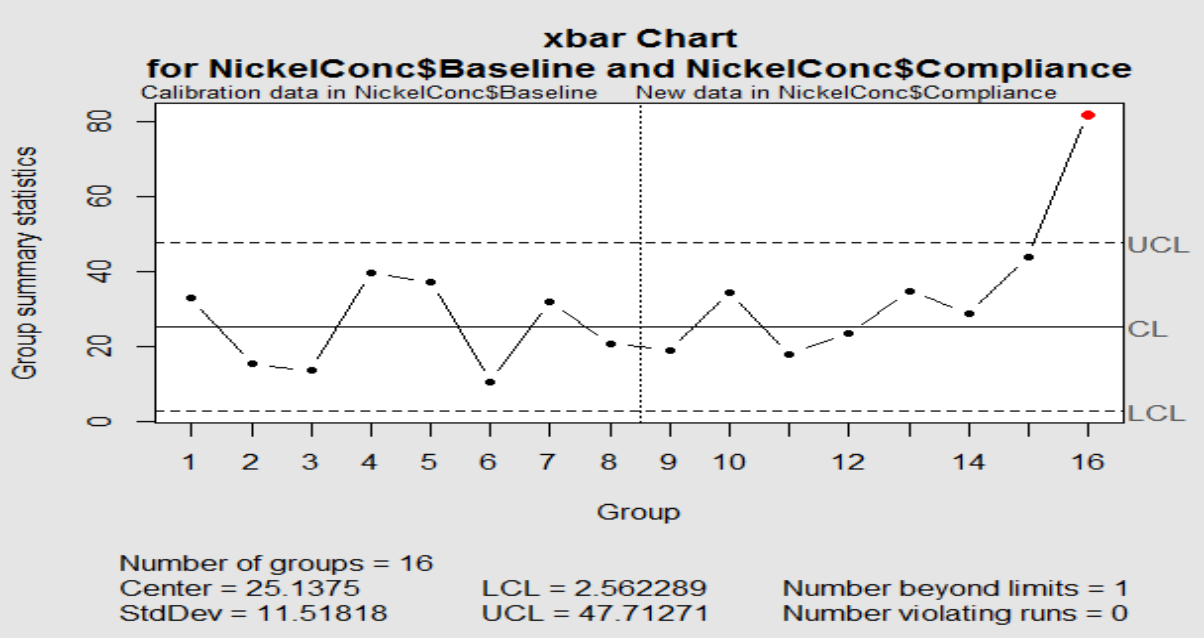
cumulative sums are greater/less than some pre-specified upper/lower decision bound, called the ***decision interval*** (recommended value: between 4 and 5).

Example 3.4 Shewhart and CUSUM control charts for Nickel Concentrations

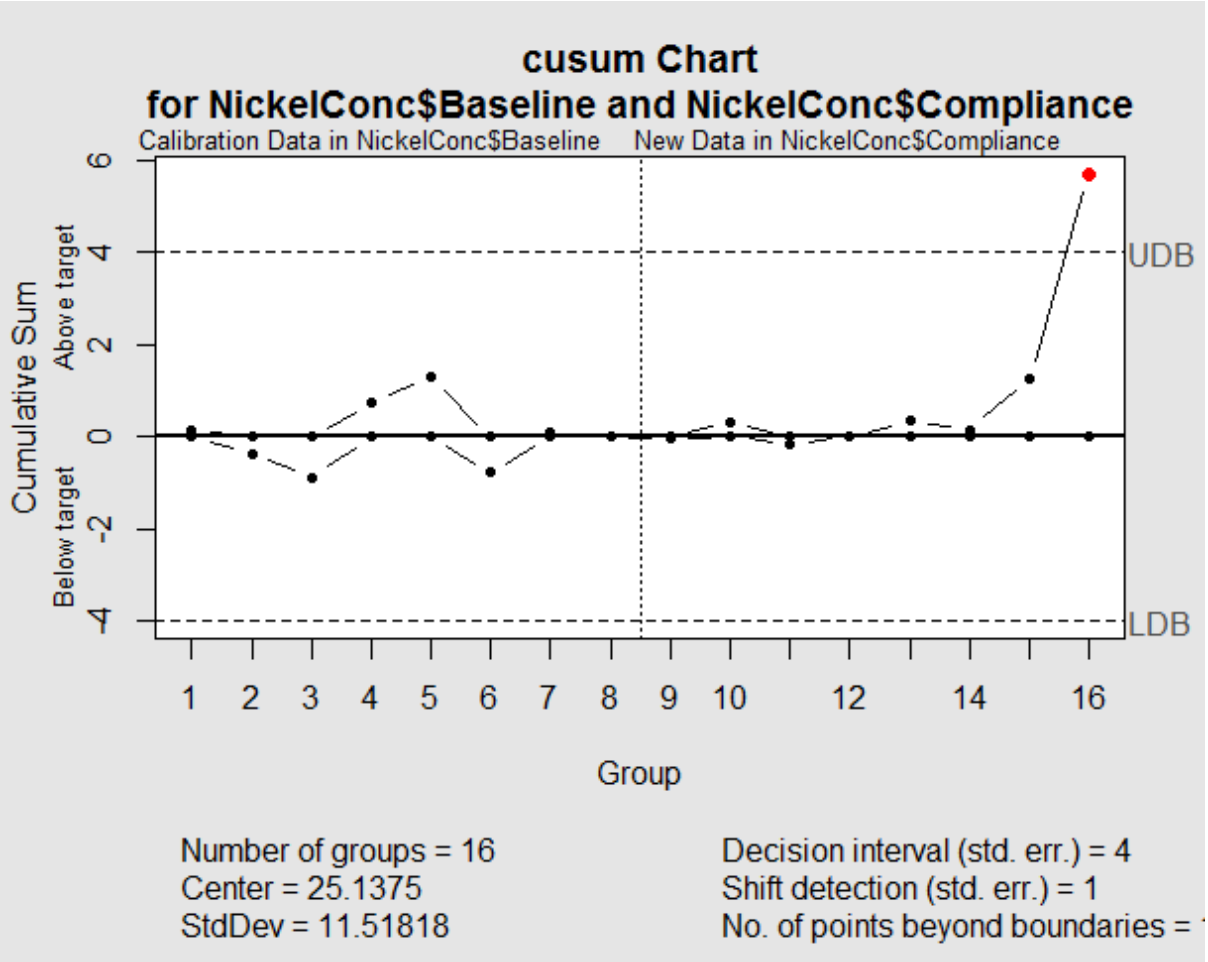
Nickel concentrations (ppb) in groundwater sampled at a compliance well over two separate 8-month periods in 1995 and 1996 are shown below (data in "*NickelConc.csv*"). The first year (1995) is considered the baseline period, and the second year is considered the compliance period. There is no evidence that these data deviate grossly from a normal distribution. The sample mean and standard deviation based on the baseline data are 25.1 and 11.5.

Month	Baseline (1995)	Compliance (1996)
1	32.8	19
2	15.2	34.5
3	13.5	17.8
4	39.6	23.6
5	37.1	34.8
6	10.4	28.8
7	31.9	43.7
8	20.6	81.8

"R": Shewhart control chart
qcc(NickelConc$Baseline, type="xbar", std.dev=sd(NickelConc$Baseline), newdata= NickelConc$Compliance, nsigmas = 3,confidence.level=0.95)



"R": CUSUM control chart
cusum(NickelConc$Baseline, std.dev= sd(NickelConc$Baseline), decision.interval = 4, se.shift = 1, newdata= NickelConc$Compliance)



Based on the two charts, we conclude that Nickel concentration was "out of control" in later 1996 at the well. ∎

Exercises

3.1 Re Exercises 1.2 and 2.1, choose two monitoring plans related to (1) hypothesis testing and (2) time series analysis from your list. Describe the monitoring processes and how to collect appropriate data sets.

3.2 Re-do the Examples in this talk.

References

- Millard, S.P. and Neerchal, N. K. (2000), *Environmental Statistics with S-PLUS*, Chapman & Hall.
- http://finzi.psych.upenn.edu/library/EnvStats/
- http://finzi.psych.upenn.edu/library/qcc/