Trend Example Applications

These example applications for S-PLUS are designed to demonstrate the use of the Kendall family of functions in the USGS library. A general knowledge of S-PLUS is assumed. For users needing introductions to S-PLUS, several are available in bookstores and online. Exmaples include Burns, Patrick, 2008, S Poetry, last accessed March 19, 2009, at *http://www.burns-stat.com/pages/spoetry.html*, and Lam, Longhow, 2001, An introduction to S-PLUS for Windows: Amsterdam, The Netherlands, CANdiensten, 230 p., last accessed March 26, 2010, at *http://www.splusbook.com*. A detailed introduction to the S language is Kraus, A., and Olson, M., 2002, The basics of S-PLUS: New York, Springer, 420 p. Another example of an introduction that concentrates on statistical applications is Venables, W.N., and Ripley, B.D., 2002, Modern applied statistics with S:,New York, Springer, 495 p.

A few type face conventions are used to assist the user in interpreting the guidelines. S-PLUS object names and function arguments are in *italics* and function names are followed by parentheses *()*, window names are highlighted in **bold**, and object class names are in a sans-serif font. User commands, called calls, are in `plain text` and on separate lines. Any printed output from a call, whether in a **Commands Window** or **Report Window**, is shown in `plain text` in a box. Plots are shown as is, copied from the default settings for an S-PLUS **Graph Sheet**; however, individual user preferences can change the way the plots are displayed. Any S-PLUS dialog menu directions are shown in **bold** separated by a vertical bar (|). Column names and other dialog box entries are underlined.

An important way to manage data is to use the **Object Explorer Window**. The **Object Explorer Window** can be opened by clicking on the small icon that is a little yellow folder with two blue dots above it  .

Introduction *Effects of ties on Kendall's tau and slope estimates*
Application 1. *Estimate the trend of an annual series*
Application 2. *Estimate the trend of a seasonal series*

Effects of ties on Kendall's tau and slope estimates

A large number of ties can cause misleading results when calculating and comparing Kendall's tau and the Sen slope estimator. As Helsel and Hirsch (2002) describe in section 8.2.1 (p. 213), tau is a function of the number of positive (concordant) pairs minus the number of negative (discordant) pairs as defined by equations 8.1 and 8.2. If ties are present, then the correction for ties in section 8.2.3 of Helsel and Hirsch (2002, p. 215) is used. In the presence of ties, tau in effect looks at just the pairs that are not tied so that tau will be zero if and only if the number of concordant pairs equals the number of discordant pairs. That is, ties—no matter how many—cannot cause a zero value of tau. Furthermore, if tau is zero, the p-value (significance level) will be exactly one (equation 8.3).

The number of tied pairs does have an effect on the p-value. This arises in the revised calculation of the variance of S (test statistic) as shown in equation 8.4. As the number of ties INCREASES, the variance of S DECREASES with the result that a given value of tau will be MORE significant.

The calculation of the slope (called the Kendall-Theil Robust Line and Theil slope estimate by Helsel and Hirsch (2002) on page 266 and also called the Sen slope estimate) is a completely separate computation from that of Kendall's tau. It is the median of all possible pairwise slopes. If there are many ties, then there will be many zero pairwise slopes. This can give rise to a zero slope estimate even if tau is nonzero and has a significant p-value! This is an unfortunate consequence when many ties are present.

Consider a time series of n (>3) equally spaced observations; the first n-1 of which are the value a, and the nth observation is the value b (>a). The calculations (using the notation of Helsel and Hirsch, 2002) proceed as

```
P = n-1
M = 0
S = P - M = n-1
tau = S / [n(n-1)/2] = 2/n
var(S) = [n(n-1)(2n+5) - (n-1)(n-2)(2n+3)] / 18
Sen slope = median{(n-1)(n-2)/2 occurrences of 0 and n-1 non-0 occurrences} = 0
For n=10, tau=0.20 with p-value=0.164 and slope=0.0.
For n=20, tau=0.10 with p-value=0.112 and slope=0.0.
```

```
For n=50, tau=0.04 with p-value=0.096 and slope=0.0.
```

The extreme number of ties has rendered the results misleading at best.

REFERENCE:

Helsel, D.R. and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey, Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p.

Estimate the trend of an annual series

This application illustrates how to estimate the trend of an annual series. The method is appropriate for annual summary statistics or individual values, such as peak flows. The trend test is described in Helsel and Hirsch (2002). The serial test is defined in Dufour (1981).

Step 1. Create the dataset

This example will use the *Flint* peak-flow dataset. The *Flint* data are the 71 annual peak flows on the Flint River near Griffin, Georgia, from 1937 through 2007. It can be created in the user's chapter by copying the following call to the **Commands Window**.

```
Flint <- data.frame(peaks=c(2980, 5200, 3650, 2080, 1040, 13000, 6310, 4170,
 7750, 9350, 5490, 4170, 13200, 875, 920, 6480, 2860, 2640, 2370, 4590, 5330,
 4730, 5180, 4730, 11100, 3900, 4850, 5810, 2880, 7520, 3880, 4550, 4170, 6820,
 12300, 7850, 5540, 3110, 7470, 9240, 3390, 5420, 6380, 4620, 5320, 4430, 4530,
 3000, 3280, 2080, 4120, 3000, 3390, 11500, 2690, 2170, 6120, 31500, 8000, 9200,
 8400, 9610, 1550, 1760, 4070, 1400, 6920, 6880, 12500, 1470, 2050),
 WY=1937:2007)
```

Step 2. Perform the trend test

The *kensen()* function performs the Mann-Kendall test and computes an estimate of slope. It is executed for these data by typing the following call in the **Commands Window**:

```
kensen(Flint$peaks, Flint$WY)
```

This call produces the following output:

```
      Kendall's tau with the Sen slope estimator

data:  Flint$peaks and Flint$WY
tau = 0.0016, p-value = 0.9881
alternative hypothesis:  slope is not equal to 0
```

```
sample estimates:
 slope median.data median.time
     0        4620        1972
```

In this case, the null hypothesis is not rejected and we assume that there is no evidence for a trend in the peak flows in the Flint River near Griffin, Georgia.

Step 3. Test for serial correlation

If the p-value from the trend test is less than a predefined critical value, which typically is 0.05, then a test for serial correlation should be done. Helsel and Hirsch (2002) state that there must be no serial correlation for the p-values to be correct in the Mann-Kendall test. The *serial.test()* function will perform either of two nonparametric tests for serial correlation. See the documentation in the USGS Help file for details. It is executed for these data by typing the following call in the **Commands Window**:

```
serial.test(Flint$peaks)
```

This call produces the following output:

```
      wilcoxon test for serial dependence

data:  Flint$peaks
S = 1483, p-value = 0.9194
alternative hypothesis:  lag-1 serial dependence is not equal to 0
```

The p-value of 0.9194 indicates that the null hypothesis should not be rejected and there is no reason to believe that serial correlation of the peaks is an issue for the trend test.

REFERENCES:

Dufour, J.M. 1981, Rank test for serial dependence: Journal of Time Series Analysis v. 2, n 3, p 117-128.

Helsel, D.R. and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey, Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p.

Estimate the trend of a seasonal series

This application illustrates how to estimate the trend of a seasonal series. A seasonal series is defined as data collected at fairly regular intervals over a period of time. The method is appropriate for any regularly spaced data. The data must be forced into a regular series, with a single value for each period and a missing value when no sample is available. The trend test is described in Helsel and Hirsch (2002).

Step 1. Create the regular series

This example will use the bicarbonate data from the *WI1majorIons.df* dataset available in the USGS library. The *WI1majorIons.df* data include 6 samples from water year 1986 and 31 samples from water year 1993 through 1995. Copy the dataset to the user's chapter by clicking and dragging the data in the left pane of the **Object Explorer Window**, or by typing the following call in the **Commands Window**.

```
WI1MajorIons.df <- WI1MajorIons.df
```

The user can display the dataset in the **Data Window** by clicking on the icon in the right pane of the **Object Explorer**. A quick examination of the data indicates that the first sample in the series of water year 1993 through 1995 was in April of 1993 and the last was in July of 1995. Because there were no samples in 1992, the regular series can begin in January 1993.

A regular series of the bicarbonate data can be created using the *regularSeries()* function in the USGS library. There are many combinations of arguments to *regularSeries ()* that can be used to construct a wide range of regular series data. If the *period* argument is set to "months" in the call to *regularSeries()* then a monthly series from April 1993 through July 1995 would be created. But the seasonal Kendall test function requires complete years. Thus, the begyear, endyear, monthlist, and daylist arguments must be set. The following call, entered in the **Commands Window**, will create an S-PLUS object called *HCO3.series*. It is a simple vector of length 36. Note that the end day of every month must be specified in the call.

```
HCO3.series <- regularSeries(WI1MajorIons.df$HCO3, WI1MajorIons.df$sample.dt,
      begyear=1993, endyear=1995, monthlist=month.abb,
      daylist=c(31,28,31,30,31,30,31,31,30,31,30,31))$x
```

Printing the data as a matrix of 12 rows shows the data that will be compared within each season. Note that no comparisons are possible in February, October, or December, month numbers 2, 10, and 12. To see the data as a matrix, enter the following call in the **Commands Window**.

```
print(matrix(HCO3.series, nrow=12))
```

```
      [,1] [,2] [,3]
 [1,]   NA  116  121
 [2,]   NA  124   NA
 [3,]   NA  127  106
 [4,]  130  120  116
 [5,]  110  131   90
 [6,]  103  112   NA
 [7,]  146   98  129
 [8,]  145  124   NA
 [9,]  142  115   NA
[10,]  118   NA   NA
[11,]  118  127   NA
[12,]  124   NA   NA
```

Step 2. Perform the trend test

The *seaken()* function computes the seasonal Kendall trend test—it requires two arguments, the series and the number of seasons per year. Type the following call in the **Commands Window** to perform the test and view the results.

```
seaken(HCO3.series, 12)
```

```
      Seasonal Kendall with correlation correction

data:  HCO3.series (3 years and 12 seasons)
tau = -0.3333, p-value = 0.332
alternative hypothesis:  slope is not equal to 0
sample estimates:
 slope median.data median.time
  -8.5       120.5         1.5
```

The p-value that is printed is selected based on the number of years. If the number of years is less than 10, then the p-value that is not corrected for serial correlation is printed. Otherwise the p-value that is corrected for serial correlation is printed. See the documentation for *seaken()* for more details.

REFERENCE:

Helsel, D.R. and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey, Techniques of Water-Resources Investigations book 4, chap. A3, 522 p.