MRI REPORT

METHODOLOGIES FOR DETERMINING TRENDS IN WATER QUALITY DATA

by

Karin M. Bauer
William D. Glauz
Jairus D. Flora

FINAL REPORT
July 3, 1984

EPA Contract No. 68-02-3938, Assignment No. 29
MRI Project No. 8205-S(29)

Prepared for

Industrial Environmental Research Laboratories
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

Attn: Ms. Susan Svirsky
Task Manager (WH-553)

METHODOLOGIES FOR DETERMINING TRENDS IN WATER QUALITY DATA

by

Karin M. Bauer
William D. Glauz
Jairus D. Flora

FINAL REPORT
July 3, 1984

EPA Contract No. 68-02-3938, Assignment No. 29
MRI Project No. 8205-S(29)

Prepared for

Industrial Environmental Research Laboratories
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

Attn: Ms. Susan Svirsky
Task Manager (WH-553)

PREFACE

Section 305(b) of the U.S. Clean Water Act requires that the States report biennially on the quality of their navigable waters. To assist the States in the preparation of these reports, the Environmental Protection Agency issues guidance information. This report on water quality trend determination was prepared by Midwest Research Institute for use by EPA as a portion of the guidance information for helping the States in preparing their 1984 reports. The authors are indebted to the several EPA reviewers who made many thoughtful and worthwhile suggestions to improve the earlier draft.

TABLE OF CONTENTS

		Page
I.	Introduction	1
II.	Purpose	4
III.	Important Considerations	6
	A. What Is A Trend?	6
	B. What Is A Change?	6
	C. Seasonal Effects	7
	D. What Comprises An Observed Series?	7
	E. How Much Data?	10
	F. How Often?	11
	G. Hypothesis Testing	12
	H. Estimation	14
	I. The Normal Distribution.	15
	J. Parametric or Distribution-Free?	20
	K. Plotting	22
	L. Organization of this Document.	23
IV.	Parametric Procedures.	27
	A. Methods to Deseasonalize Data	27
	B. Regression Analysis.	28
	C. Student's T-Test	32
	D. Trend and Change	38
	E. Time Series Analysis	39
V.	Distribution-Free Methods.	41
	A. Runs Tests for Randomness.	41
	B. Kendall's Tau Test	45
	C. The Wilcoxon Rank Sum Test (Step Trend).	50
	D. Seasonal Kendall's Test for Trend.	57
	E. Aligned Rank Sum Test for Seasonal Data (Step Trend)	64
	F. Trend and Change	69
VI.	Special Problems	71
	A. Missing Data	71
	B. Outlying Observations.	72
	C. Test for Normality	73
	D. Detection Limits	75
	E. Flow Adjustments	76
Bibliography		78
Appendix		82

List of Figures

Figure	Title	Page
1	Composite Series.	8
2	Sample Trend Line	9
3	Two Normal Distributions of a Variable X.	16
4	Standard Normal Distribution.	17
5	Decision Tree	24
6	Sample Linear Regression Line	29
7	Monthly Concentrations of Total Phosphorus.	58
8	Plot of Percentage Violations Versus Time	61
9	Example of Plot on Probability Paper.	74

List of Tables

Table	Title	Page
1	Upper Tail Probabilities for the Standard Normal Distribution.	19
2	Two-Tailed Significance Levels of Student's T	33
3	The 5% (Roman Type) and 1% (Boldface Type) Points for the Distribution of F	37
4	r-Tables Showing 5% Levels for Runs Test.	42
5	Upper Tail Probabilities for the Null Distribution of Kendall's K Statistic.	48
6	Upper Tail Probabilities for Wilcoxon's Rank Sum W Statistic	53

SECTION I. INTRODUCTION

Water quality reports prepared by the States and jurisdictions under Section 305(b) of the Clean Water Act contain information on a wide variety of parameters. Because the reports cover a relatively short 2-year period, it is often difficult to determine with reasonable certainty whether water quality has improved, remained constant, or degraded from one reporting period to the next. In order to assess water quality trends, methods must be used which differentiate a consistent trend in a given direction from natural cyclic or seasonal changes or occasional excursions from the norm.

Generally, water quality data are affected by seasonal or cyclic effects, an episodic or regular effect, a long-term monotonic trend, and random noise. The purpose of this document is to provide a guidance in the usage of statistical methods to separate out these effects and to test for each of them.

The statistical methods described in this document provide the means for the analyst to make this consistent trend assessment by using and testing concepts of data validity and significance. These statistical techniques will allow the analyst to detect the presence of long-term trends and to put confidence intervals on the magnitude of trends or changes.

Once a specific statistical approach is established, it can be used in succeeding years. As a result, the data-gathering process itself may be modified, thereby increasing the quality and value of future data. Another potential benefit is the discovery that historical data can have previously unappreciated value; retrospective analyses then become possible.

In principle, statistical methods can be applied to all the data collected during the water quality monitoring process. These data include chemical parameter values for water, sediment, and tissue samples (e.g., DO, pH, nutrients, trace metals, specific organic compounds); physical parameter values (e.g., turbidity, suspended solids, light penetration); biological parameter values (e.g., macroinvertebrate populations, other aquatic species); pollutant discharge or source data; and derived values which combine several parameters in a composite evaluation of water quality (e.g., trophic indices, water quality indices, violation statistics). However, the water quality analyst is advised to apply statistical methods initially to only those parameters which are of greatest interest or priority and which offer the most extensive and reliable data. As the analyst becomes more familiar with statistical methods, he will then be able to expand the scope of analysis to most or all of the parameters.

Suggestions for the scope of an initial effort to use statistical methods in determining water quality trends are as follows:

1. The analyst *should* select for analysis those chemical water quality parameters which are of greatest importance to the State, to selected areas of the State, or to the designated use in question.

2. The analyst should consider important biological parameters such as macroinvertebrate populations and distributions.

3. A minimum of two or three of the important parameters, including at least one from each of the categories defined in (1) and (2), should be statistically analyzed by methods designed to obtain valid estimates of trends.

4. The analyst will probably find that the period of these data bases will need to be extended beyond the 2-year period specifically called for in the 305(b) reports, so parameters should be selected for which comparable data were collected in earlier years. Data from previous 305(b) reporting periods could be used to augment the current period data.

These analyses should be conducted first for specific water bodies rather than statewide. Further, it may be necessary to initially limit an analysis to one part of a specific water system, with later extension to the remainder of the system. The results of data analysis should tell the analyst which is the appropriate and valid scope.

As mentioned earlier, the analyst may find it appropriate to use a statistical approach to modify ongoing data-gathering activities. In addition to improving the quality of the water quality data which are collected, this may reduce the cost of data collection because data will now be acquired more selectively.

In summary, there are two basic perspectives from which the analyst should view the results of the application of statistical methods to water quality data: the importance of the detected trends themselves, for what they say about both the State's water quality and its water quality programs; and the impact of the statistical methods and their results on ongoing water quality data collection activities.

SECTION II. PURPOSE

The purpose of this document is to provide guidance to the States and Regions in analyzing for trends in their water quality data. It is not a statistical treatise. Indeed, some statisticians may be concerned at the looseness with which terms and symbols seem to be used; the lack of mention of all the ifs, ands, and buts; and the possibility that the reader may end up indiscriminantly using one test when another is more proper or more powerful.

Rather, this document is aimed at persons with little background in statistics, or even much in algebra. Most of the methods discussed require only simple arithmetic and the use of standard tables. As such, it is a "how to" approach, with little or no theory. In that respect, it is rather like the statistical programs available now on most computers, which are designed so that a nonstatistician can readily feed in data, select the statistical method to be used, and then read and interpret the answers. In fact, many of the methods are available as computer programs, and examples of these are given in this document.

To be sure, there are distinct dangers in this approach, just as there are when nonstatisticians use the "canned" computer programs. It is always possible to use an inappropriate or incorrect test. This document will therefore warn the reader of the major limitations and potential pitfalls to be avoided. Ideally, of course, the analyst would always select the "best" test; however, the "best" test is likely to require extra effort, access to special computer programs, and the like. Realizing this, the analyst may be tempted to give up and do nothing at all, or, worse, make arbitrary judgments about "trends" in his data, with no justification. It seems better, in such cases, to have some statistical backup, even if it is limited, before such determinations are made.

Of course, if the analyst does have some background in statistics, has a statistician available to assist or as a consultant, or has access to statistical computer programs, these advantages should be applied. In such cases, the material in this document should assist the analyst in making even better use of his available resources.

Finally, for the reader desiring further information, all the descriptions of the methods and tests given here include references to the (more rigorous) statistical literature.

SECTION III. IMPORTANT CONSIDERATIONS

A. What Is A Trend?

The concept of "trend" is difficult to define. Generally, one thinks of it as a smooth, long-term movement in an ordered series of measurements over a "long" period of time. However, the term "long" is rather arbitrary and what is long for one purpose may be short for another. For example, a systematic movement in climatic conditions over a century would be regarded as a trend for most purposes, but might be part of an oscillatory or cyclical movement taking place over geological periods of time. In speaking of a trend, therefore, one must bear in mind the length of the time period to which the statement refers. For our purposes, we shall define a trend to be that aspect of a series of observations that exhibits a steady increase or decrease over the length of time of the observations, as opposed to a "change," described next.

B. What Is A Change?

It is important to distinguish between the terms "trend" and "change"; they are often incorrectly used interchangeably in water quality reports. However, they are not equivalent terms and should be carefully distinguished. A "change" is a sudden difference in water quality associated with a discrete event. For example, suppose for a period of time at a given station, concentrations of some pollutants were persistently high. Then a new treatment facility is placed into operation and, from then on, the concentrations of these same pollutants are lower. This sudden improvement in stream water quality should be referred to as a change, not a trend. A change is sometimes called a "step trend" and, for clarity, the other is called a "long-term trend."

C. Seasonal Effects

Observations taken over time at regular intervals often exhibit a "seasonal" effect. By this, a regular cycle is meant. For monthly data a seasonal effect is a regular change over the year that more or less repeats itself in succeeding years. This might be associated with different flow rates corresponding to the seasonal pattern of precipitation and melting in the watershed above a particular observation station. Here, seasonal components will be restricted to refer to annual cycles, although in other applications with different frequencies of observation, cycles could occur quarterly, monthly or daily.

D. What Comprises An Observed Series?

A series of observations can be broken into several parts or components, as illustrated in Figure 1 below. In this figure, each component of the series has been generated separately. The seasonal component is represented by a sine wave plotted in part a. A discrete change occurred at month 18 and is shown in part b, along with a linear trend represented by the straight line. Finally, part c shows an irregular or random component. These parts are summed to give the series as it would be observed in part d of Figure 1. In practice, one observes a series of data such as illustrated in part d of Figure 1. The objective of trend analysis is to determine whether such a series has a significant long-term trend or a discrete change associated with a particular event.

It is often the case that a trend is not immediately obvious from a series of "raw" measurements. Figure 2 depicts a series of ammonia concentration values measured monthly at a single station. Statistical techniques can be used to obtain the "deseasonalized" data (more will be said about this later), and the trend line for the deseasonalized data. (This particular example was taken from STORET). The horizontal axis represents time (months) and the vertical axis is the ammonia concentration in mg/l. Trend analysis techniques attempt to sort out the data displayed in Figure 2 into components such as those illustrated in Figure 1 using statistical

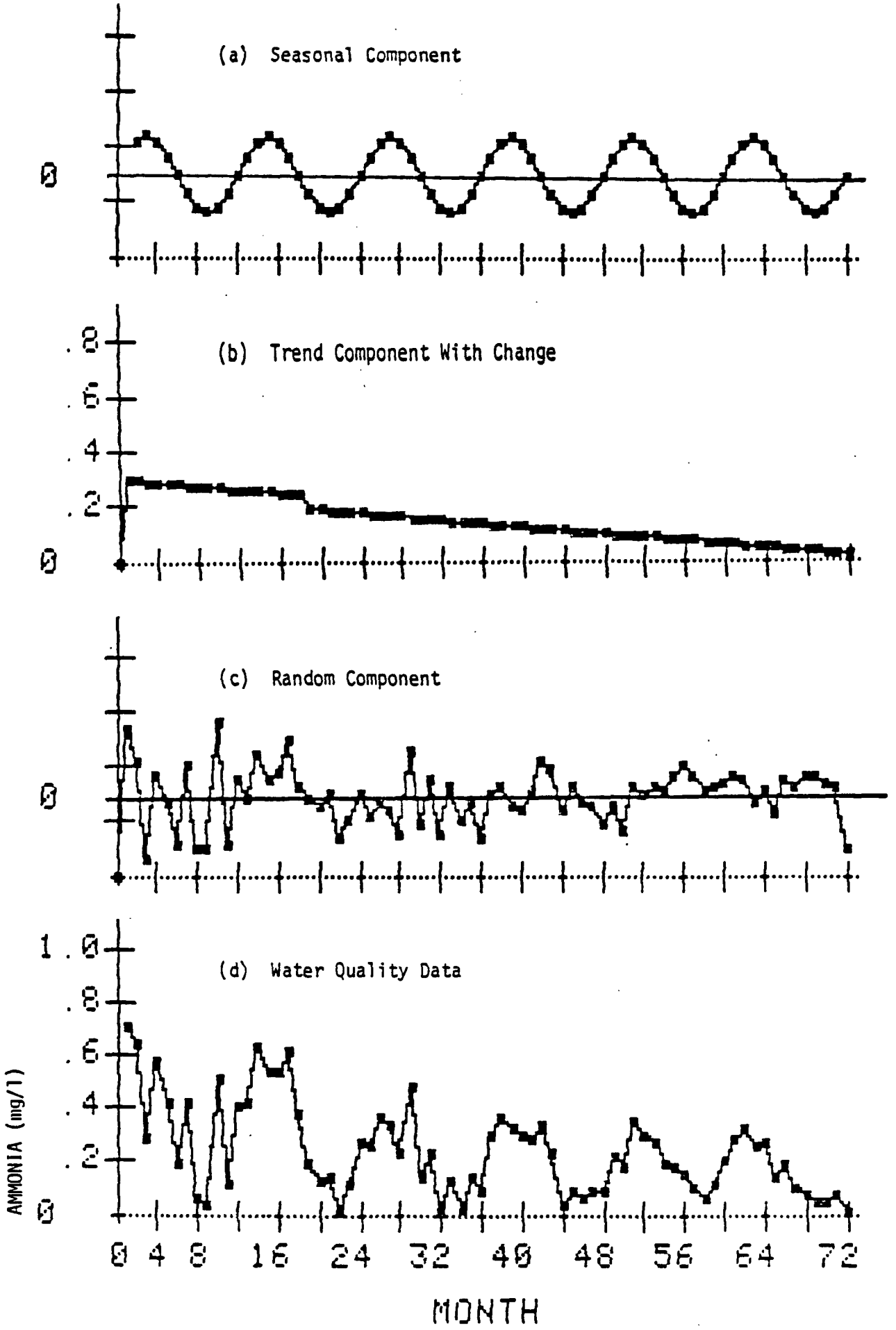


Figure 1. Composite Series

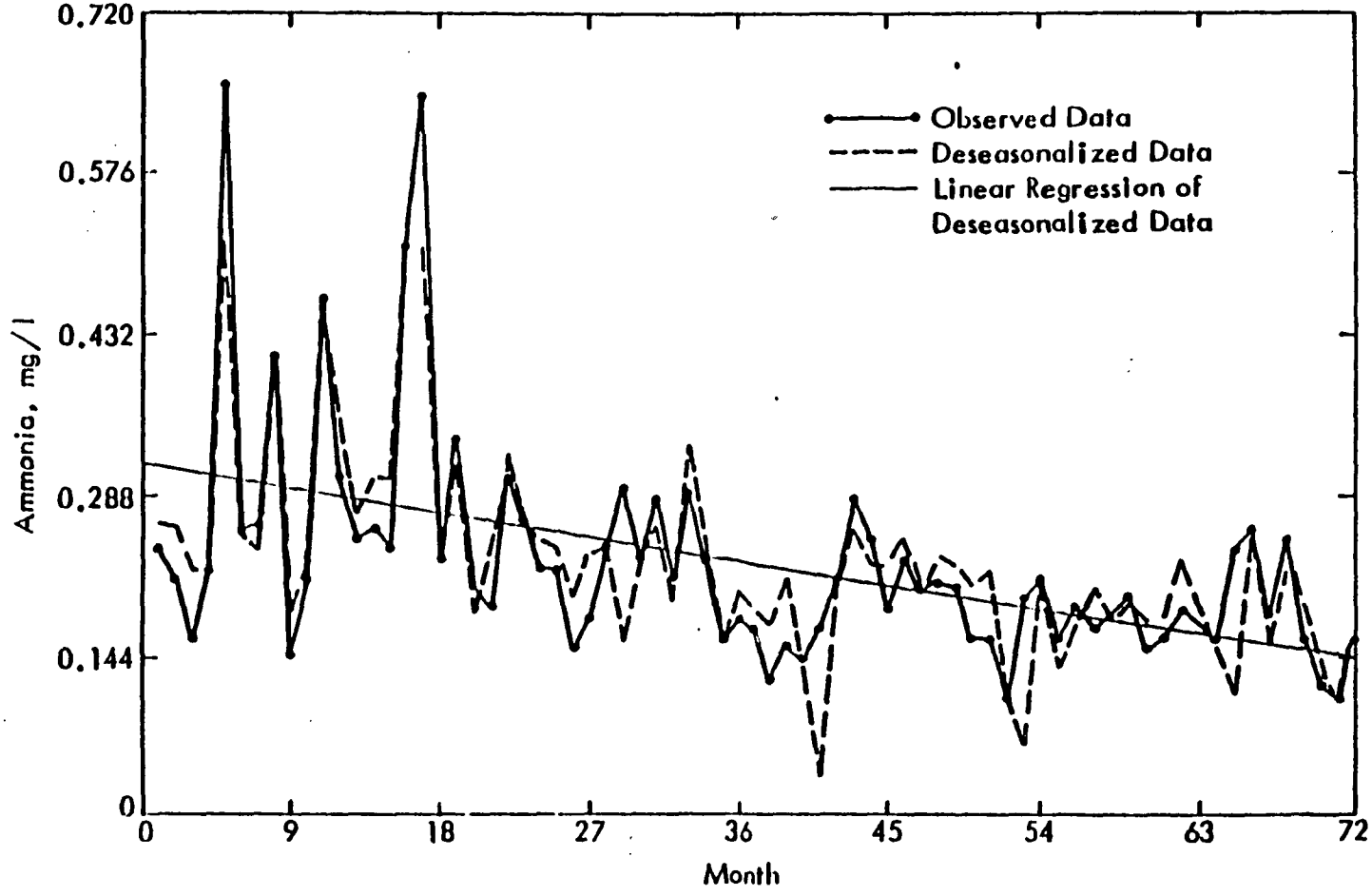


Figure 2. Sample trend line.

techniques to test for the presence of a trend while adjusting for a seasonal effect. The techniques discussed as trend analysis provide methods to determine whether the data show a trend, a change (often referred to as a step trend), its direction, and whether such a change or trend is statistically significant--larger than can be ascribed to random variation.

Most of the methods presented later in this document deal with trends, although some tests dealing with changes are also included. We concentrate on simple techniques. More detailed analyses of such data are referred to as time series analysis and can become quite complex. These more detailed methods are discussed briefly in Section IV-E.

E. How Much Data?

Three or four measurements do not make a trend. Imagine a friend flipping a coin and getting "heads" four times in a row. We might consider that to be an interesting result, but not one that would prompt us to conclude that the coin had heads on both sides. The amount of data is just not convincing enough. However, if the friend continued flipping the coin and obtained 10 heads in a row, we would then seriously suspect that the coin was indeed unusual. This assumes, of course, that we can rule out the possibility that the friend is somehow manipulating the outcome of the coin tosses and thus biasing the results.

The amount of water quality data needed depends on the frequency of collection and the period of seasonality considered. Theoretically, the amount of data needed can be quantified using the laws of probability if the specific requirements of the analysis are stated. Such detailed sample size calculations are quite specific to each situation. However, a general discussion of the considerations can give some guidance. Referring to part a of Figure 1, one can see that a short series of six to eight monthly observations would give a rather misleading impression of the series. It could show a marked increasing trend, a decreasing one, or one of two types of curvature. Clearly, to identify a cyclical pattern, the number of

observations must be large enough to cover two complete cycles and preferably more. Thus, with monthly data and a seasonal effect, a minimum of 24 to 30 months of data would be needed. On the other hand, if data are aggregated to an annual basis, one might be able to identify a trend based on as few as 5 or 6 years. However, such an identification would not be able to distinguish between a discrete change and a long-term trend, nor whether the observed trend might be part of a longer cycle.

The more data, the more components of a series could be identified. One rule of thumb is that there should be 10 observations (some authors prefer 20) for each component to be tested for or adjusted for. In addition, for cyclical effects, the series should cover at least two full periods. The 2 years of data reported in 305(b) would be just barely sufficient to apply a technique that includes seasonality. It would be preferable to include data from previous reporting periods, provided that they are compatible.

F. How Often?

The sampling scheme, i.e., the frequency of data collection, and the number of sites and parameters, dictates to a great extent the type of statistical procedures which should be used to analyze the data. Most water quality characteristics (physical, chemical, biological) are collected on a monthly basis; others are obtained once a year.

Examples of water quality characteristics measured in the EPA's basic water monitoring program include:

Characteristic	Sampling frequency in rivers and streams
Flow	monthly
Temperature	monthly
Dissolved oxygen	monthly
pH	monthly
Conductivity	monthly
Fecal coliform	monthly
Total Kjeldahl nitrogen	monthly
Nitrate plus nitrite	monthly
Total Phosphorus	monthly
Chemical oxygen demand	monthly
Total suspended solids	monthly
Representative fish/shellfish tissue analysis	annually

The important point is that the trend analysis must consider sampling frequency when determining the appropriate type of trend analysis. If a cyclical component is suspected, the frequency must be a relatively small fraction of the period in order to estimate the cycle. In addition, data for at least two cycles would be needed. On the other hand, if data are collected at intervals as long as or longer than a possible cycle, it is important to ensure that data collection times are at the same point in the cycle. Otherwise, spurious trends might appear or the variability of the data might be substantially increased. Care should be taken to ensure that data collection is under the same conditons each time a sample is obtained.

G. Hypothesis Testing

In hypothesis testing, a statement (hypothesis) to be tested is specified. It is generally stated with enough detail so that probabilities of any particular sample can be calculated if it is true. The hypothesis to be tested is referred to as the "null" hypothesis. An example is the hypothesis that a set of data are independent and identically distributed according to the normal distribution with mean zero and variance

one. This would correspond to the hypothesis of no trend. (For most water quality parameters a different mean and variance would be appropriate.) A second hypothesis--the alternative--may be specified. This represents a different situation that one wishes to detect, for example, a trend or tendency for the concentration of a pollutant in water to decrease with time.

To test a hypothesis, one observes a sample of data and calculates the probability of the data assuming that the hypothesis is true. If the calculated probability is reasonably large, then the data do not contradict the null hypothesis, and it is not rejected. On the other hand, if the observed data are very unlikely under the null hypothesis, one is faced with concluding either that a very unlikely event has occurred, or that the null hypothesis is wrong. Generally, if the observed data are less likely than a prespecified level, one agrees to conclude that the null hypothesis is wrong and to reject it.

In testing a hypothesis one can make two types of errors. First, one can reject the null hypothesis although it is true: this is called an error of the first kind or a Type I error. The probability of a Type I error is often denoted by the letter α. An example of a Type I error would be stating that a trend existed when the data were actually random. On the other hand, one can erroneously accept the null hypothesis when in fact it is false; this is called an error of the second kind or a Type II error. The probability of a Type II error is often denoted by the letter β. An example of a Type II error would be the failure to detect a real trend in water quality. To decide whether to reject a null hypothesis, probability levels usually expressed as percentages are used. These probability levels are also called significance levels and are commonly chosen to be 1, 5, or 10%. In addition, confidence levels are defined as 100% minus the significance levels, e.g., 99, 95, or 90%, respectively.

Consider again the friend with the coin. If the coin and the friend are unbiased, there is an equal chance that on any given toss the coin will come up heads or tails. The probability, usually denoted by p, that it will be heads is thus 50% or 1/2. The probability of getting heads four times in a row is:

$(1/2)(1/2)(1/2)(1/2) = 1/16$, or 6.25%.

That is, the odds are only 1 out of 16 that this result would occur purely by chance. If the friend repeated the experiment 16 times, only once (on the average) would he get heads 4 times in a row.

If one agrees to reject the hypothesis that $p = 1/2$ if he gets four heads in a row, then the significance level of 0.0625 is the mathematical probability of failing to accept the null hypothesis that the coin has both a head and a tail that are equally likely to come up (based on this experiment). An α-value of 0.0625 could be considered statistically "significant." If one decides to reject only if heads appeared 10 times in a row, the probability of the friend getting heads 10 times in a row is 1/2 multiplied by itself 10 times, which is approximately 0.001, or 1 chance out of a 1,000. Such a finding would usually be considered "highly" significant, indeed.

H. Estimation

In addition to testing a hypothesis (of no trend against the alternative that a trend exists for example), one may wish to estimate the magnitude of a trend. Such a measure might be a change in concentration per year. Two aspects of statistical estimation need to be kept in mind. First, one may obtain a point estimate, which is a statistic that gives the "best" single value for the size of the trend. However, this by itself is of little use. It needs to be coupled with a test to determine whether that value is significantly different from zero. An additional formulation is to give a range or an interval estimate for the value in question. Such an interval in statistics is called a confidence interval. This is an interval calculated from the data in such a manner that it would contain the true but unknown value of the measure in a specified proportion of the samples. The proportion of the samples that would yield an interval that contains the

measure is called the confidence level. As mentioned previously, it relates to the significance level, α, in that it is 100% minus the significance level.

In general, one would want to estimate the magnitude of any trend and to place confidence limits on the magnitude of the trend. If the sample size is small, and/or the variance is quite large, these confidence limits will be very wide, indicating that the trend is not well estimated. If the confidence limits are quite close, one can be confident that the trend is well estimated.

I. The Normal Distribution

The results of performing an experiment such as flipping a coin or measuring the pH of a water sample can be tabulated as it is repeated over and over again. If the frequency of each observed result is then plotted against the result itself, the graph is called the frequency distribution, or distribution for short. A very important type of distribution often used in conjunction with trend analysis (and many other types of statistics) is called the normal distribution.

The normal distribution (also called the Gaussian distribution) has dominated statistical practice and theory for centuries. It has been extensively and accurately tabulated, which makes it convenient to use when applicable. Many variables such as heights of men, lengths of ears of corn, and weights of fish are approximately normally distributed. In some cases where the underlying distribution is not normal, a transformation of the scale of measurement may induce approximate normality. Such transformations as the square root and the logarithm of the variable are commonly used. The normal distribution may then be applicable to the transformed values, even if it is not applicable to the original data.

In many practical situations the variable of interest is not the measurement itself (e.g., the weight of a fish), but the average of the measurements (e.g., the average weight of 100 fish in a survey). Even if the distribution of the original variable is far from normal, the central-limit theorem states that the distribution of sample averages tends to become normal as the sample size increases. This is perhaps the single most important reason for the use of the normal distribution.

The normal distribution is completely determined by two quantities: its average or mean (μ) and its variance (σ^2). The square root of σ^2,(σ), is called the standard deviation. The mean locates the center of the distribution and the standard deviation measures the spread or variation of the individual measurements. Figure 3 depicts two normal distributions, both with mean 0, but with different standard deviations.

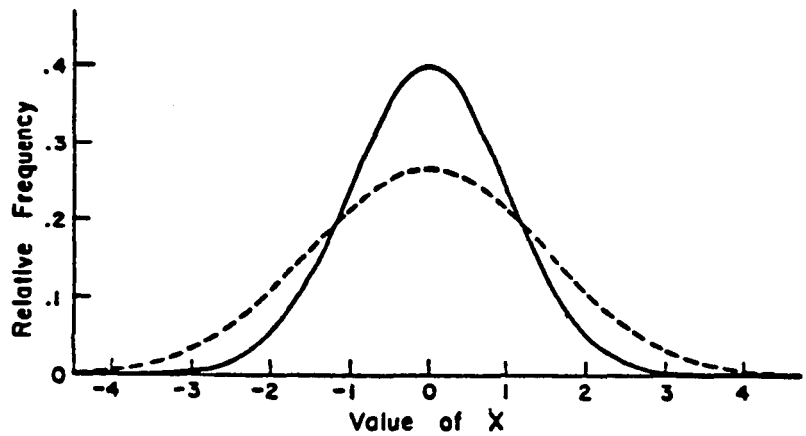


Figure 3. Two normal distributions of a variable X.
Solid line: $\mu = 0$, $\sigma = 1$
Dotted line: $\mu = 0$, $\sigma = 1.5$

The values of the mean and variance of water quality parameters could be practically anything, depending upon the parameter, when and where it was measured, etc. In order to simplify calculations it is desirable to change (transform) the data to obtain a mean of 0 and a variance of 1. To do so, compute a new variable

$$Z = \frac{X-\mu}{\sigma}$$

where X is the original variable with mean μ and variance σ^2. Then Z has a mean of 0 and a variance of 1. The quantity Z is called the standard normal deviate, and has what is known as the standard normal distribution, which is extensively tabulated. (From Z, X can be computed as $X = \sigma Z + \mu$.)

In statistical testing one compares a computed value with values in a table. There are basically two types of tests--one-sided and two-sided. In the coin flipping example where we were interested only in the result involving all heads, one would use a one-sided test. However, if the whole argument were repeated without reference to heads specifically, but only to the fact that the same outcome was achieved four times or ten times in a row, the odds would all change. The chances of getting four like results in a row is only (1/2)(1/2)(1/2)=1/8, because the first flip is immaterial; it only matters that the last three match the first one, whatever it was. In this case one would use a two-sided test. One often refers to the one-sided test as using one tail (end) of a distribution (see Figure 4 below), and the two-sided test as looking at both tails of the distribution. (By the way, the term, two-sided, really has nothing to do with the two sides of a coin, nor does a tail of the distribution refer to the tail side of a coin.)

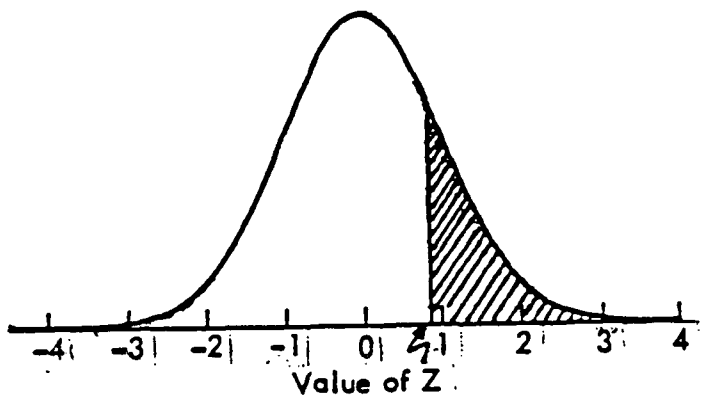


Figure 4. Standard normal distribution
($\mu = 0$, $\sigma = 1$)

Table 1 below shows one way of tabulating the standard normal distribution. The following is an explanation of how to use the table for testing. Each number in the table corresponds to the shaded area in Figure 4 for a particular, positive value of Z.

The water quality analyst can use the table either for a two-sided test (e.g., is there a trend?) or for a one-sided test (e.g., is the trend increasing?). The two-sided test would normally be used but, for simplicity, the one-sided test will be explained first.

Suppose a calculated Z from some test is 1.53, that the alternative hypothesis calls for a one-sided test, and we want to know if Z is significant. We then ask, what is the probability of obtaining a Z greater than or equal to 1.53. Referring to Table 1, in the left hand column chose the line at 1.5; across the top, chose the column at 0.03, since 1.53 = 1.5 + 0.03. The number at the intersection of the line and the column is 0.0630. The probability of obtaining a Z greater than or equal to 1.53, purely by chance, is therefore 0.0630. If one had previously chosen a significance level of 5% (α = 0.05), this value of Z would not be significant because the obtained value, 0.0630, is greater than the chosen significance level of 0.05. That is we could not say confidently that the trend being tested was real; the data could well be just random numbers.

To use the table for a two-sided test, with the same value of Z, we first obtain the same probability from the table. However, the probability we have to use is that of obtaining a Z greater than or equal to 1.53, or less than or equal to -1.53. This probability is twice 0.0630 or 0.1260 since the curve is symmetrical about Z = 0.

In summary, for the calculated Z of 1.53, we determined the one-sided significance level as being 0.063 and the two-sided significance level as being 0.126. Table 1 can also be used the other way around. For a given significance level α of 0.025, for example, we wish to determine a value, Z_c,

TABLE 1

UPPER TAIL* PROBABILITIES FOR THE STANDARD NORMAL DISTRIBUTION

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
→ 1.5	.0668	.0655	.0643	(.0630)	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

*For a two-tailed significance level of x , multiply probability in table by 2.

Source: Hollander and Wolfe (1973) p. 258

such that a value of Z obtained from our data smaller than Z_c would not be significant while a value of Z equal to or greater than Z_c would be significant. Z_c is therefore called the critical value associated with the significance level α. In the case of $\alpha = 0.025$ and a one-sided test, the critical value would be 1.96 (row at 1.9 and column at 0.06). As before, the probability of Z being less than or equal to -1.96 is also 0.025. We can thus say that the probability of Z being between -1.96 and +1.96 is 1-2(0.025) = 0.95, or 95% of the values of Z are between -1.96 and +1.96. Thus, $Z_c = 1.96$ is the critical value for $\alpha = 0.025$ (one-sided) or $\alpha = 0.05$ (two-sided). If we want to determine, directly, the critical value for $\alpha = 0.025$ (two-sided), we look in Table 1 for $0.0125 = \alpha/2$ and find $Z_c = 2.24$ (row at 2.2 and column at 0.04).

J. Parametric or Distribution-Free?

Suppose there are two years of monthly water quality data and we want to test whether the water quality in one year differed from that in the other, based on the level of a certain pollutant. The classical method to test for differences in the concentration of the pollutant in the two years would be to compute a two-sample t-test. This test, like all statistical tests, is based on a number of assumptions about the underlying distribution from which the data were drawn. For the two-sample t-test, these assumptions are: (1) the errors are independent, (2) the variances are the same, (3) the distribution is normal, and (4) the null hypothesis is that the two means are the same. Procedures such as those that specify the form of the underlying distribution (e.g., normal) up to a few parameters are referred to as parametric procedures. Many of these are based on the normal distribution or on sampling distributions (e.g., t-, F-distribution) derived from it.

An alternative approach is to base the test statistic on less detailed assumptions about the underlying distribution. For example, if the null hypothesis merely specifies that the distribution of the water quality measure was the same in the two years and was continuous, then a nonparametric or distribution-free test can be used. The most widely used such test is

the Wilcoxon-Mann-Whitney two-sample rank test (see Section V-C). Because it does not depend on the assumed form (e.g., normal) of the underlying distribution, inferences based on the Wilcoxon rank sum test may be more reliable. On the other hand, a distribution-free test will typically have lower power to detect specific alternatives than will the appropriate parametric test if all the assumptions of the parametric test hold. However, in most cases the difference is negligible.

In order for some of the commonly applied parametric statistics to be valid, the data should be approximately normally distributed, or capable of being transformed so that they become so. A check for normality is discussed in Section VI. Unfortunately, most water quality data are often far from being normally distributed. There are two drawbacks to transforming the data. One is that it may be difficult to select a suitable transformation. The second is that the transformation may induce a scale that is difficult to interpret, or might change some of the other assumptions.

If the classical parametric statistical methods are not valid for the data, we must rely on the less familiar nonparametric methods. These procedures require fewer assumptions, so they have wider applicability. Most commercial statistical computer program packages such as SAS, BMDP, and SPSS include some nonparametric methods for data analysis, although these often are restricted to hypothesis testing applications. (They may indicate whether or not a trend exists, but not provide estimates or confidence intervals of its magnitude).

In addition to the assumption of normality (or another specific distribution), statistical procedures--particularly parametric procedures--are based on a number of other assumptions. The most common of these are constant variance, linearity of a trend, and independence of errors. The distribution-free procedures are less sensitive to violations of the assumption of equal variances. Further, the distribution-free procedures generally deal only with a monotonicity requirement for a trend rather than a specifically linear one, so they may be better for testing. However, procedures such as regression can easily accommodate various forms of a trend such as

polynomials in estimating the magnitude of the trend. Both types of procedures assume independence of errors and can be sensitive to correlation of errors. Correlated errors call for more complex time series analysis that specifically incorporates the error structure.

Many of the methods described in this document are nonparametric. However, some of the more common parametric tests are also presented. If the data appear to be normally distributed or can be transformed to be so, the user may prefer these more widely known tests.

K. Plotting

Before using any statistical method to test for trends, a recommended first step is to plot the data. Each point should be plotted against time using an appropriate scale (e.g., month or year). Figure 2, presented earlier, is an example of such a plot.

In general, graphs can provide highly effective illustrations of water quality trends, allowing the analyst to obtain a much greater feeling for the data. Seasonal fluctuations or sudden changes, for example, may become quite visually evident, thereby supporting the analyst in his decision of which sequence of statistical tests to use. By the same token, extreme values can be easily identified and then investigated. Plots of water quality parameters, expressed as raw concentrations or logarithms, loadings, or water quality index, can easily be plotted against time using STORET. Also, SAS features a plotting procedure, less sophisticated, however; for the reader familiar with SAS GRAPH, a series of computer graphics is available if the appropriate hardware is on hand.

Another area where plotting plays an important and instructive role is when testing for normality of data. A relatively simple plotting procedure is available to the user, in the case when the data base is not too large, since the plotting is done by hand. All one needs is probability paper. The use of probability paper is illustrated with an example in Section VI.

L. Organization of this Document

The normal-theory or parametric procedures are discussed in Section IV, while the distribution-free procedures are grouped into Section V. In Section IV, after methods to deseasonalize data are presented, a test for long-term trend is discussed, then a test for step trend (change), followed by a test for both types of trend. Finally, time series analysis is briefly discussed in the last subsection.

Section V is organized in a similar fashion. After discussing tests for randomness in general terms, distribution-free methods applicable to non-seasonal data are presented first for long-term trends and step trends (Subsection B and C, respectively). Next, tests for long-term trends and step trends in the case of seasonal data are discussed in Subsections D and E, respectively. The section closes with a brief discussion of the difficulties of dealing with the data when both types of trends are present.

Figure 5 below presents a decision tree indicating how to determine which procedures are appropriate based on the characteristics of the data available to the analyst. Thus, one can follow this as a road map and refer to the sections where the methods to answer each question on the diagram are discussed.

The first step is to determine whether the data are on an annual basis or were more frequently recorded. If they are more frequent than annual, then seasonality is an important consideration. In general, one must test for seasonality if the data are on a quarterly or monthly basis. If seasonality is found, it must be removed before a test for change or trend can be done. Tests for seasonality can be based on runs tests, turning point tests, serial correlations, or other general tests for randomness. Methods for removing seasonality, or for accounting for it in the analysis, are given in Sections IV and V, where appropriate.

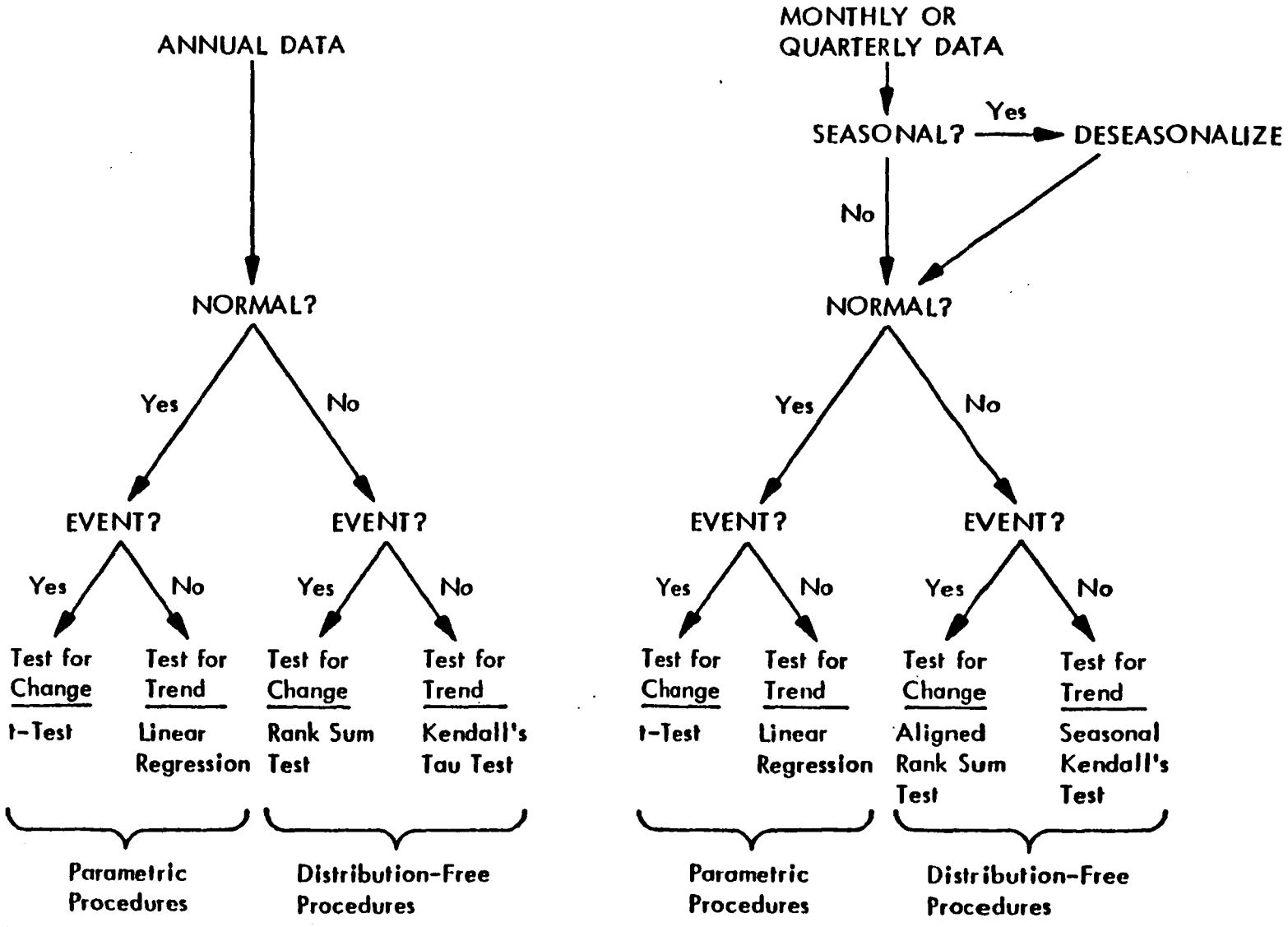


Figure 5 - Decision tree.

(parametric) procedures or distribution-free techniques. One important consideration in this decision is whether the data are normally distributed or not. If the data follow the normal distribution to a reasonable approximation, then the parametric procedures would be preferred. If the data are non-normal or there is a question, then it may be better to rely on the distribution-free procedures. Methods to assess normality are discussed in Section VI-C.

Next, it is necessary to decide whether some "event" occurred during the period covered by the data. This decision is based on external factors--for example, was a new treatment plant put into operation at a specific point in time. If it is decided that such an event did occur, the one would use an appropriate test for a change. If no discrete event occurred that could be expected to lead to a change in level of one (or more) of the water quality parameters, then one would test for a trend. This decision can take place either before or after the assumption of normality has been tested.

Testing for a change is based on the t-test if the data are reasonably normal. Detailed instructions about performing the t-test are presented in Section IV. If a distribution-free approach to testing for a change is appropriate, then the test to be used is the Wilcoxon rank sum test, assuming that there is no seasonality present in the data. Details for applying the Wilcoxon rank sum test to test for a change and to estimate the magnitude of the change are presented in Section V-C. If seasonality is present, then an aligned rank sum test must be used to adjust for the seasonality. Details of this procedure are presented in Section V-E.

If there is no event that can be identified with a likely change in level, then a test for trend is appropriate. If normal-theory procedures are deemed appropriate, then this is based on a linear regression. The use of regression to estimate and test for a trend is discussed in Section IV. If normal theory procedures are not thought appropriate, or if analysis of the residuals from the regression indicates serious violations of the assumptions, then a distribution-free procedure is called for. In the event

that there is no seasonality (as with annual data) the test procedure is based on Kendall's tau. Application of this test is presented in Section V-B. If seasonality is present, then a modified version of Kendall's test can be used. It is presented in Section V-D.

A number of special problem areas are covered in Section VI. These include how to handle missing observations; what to do with observations that appear strange (outliers); how to test for normality; how to deal with situations, sometimes fairly common, where the parameter value of interest could not be measured because it was less than the detection limit of the measurement technique; and how to make corrections to the data for variations in flow (especially important with concentration parameters).

The document concludes with a selected bibliography. The papers and texts listed are not all referenced directly, but provide additional information on one or more of the tests discussed here. For general coverage of statistical testing, the texts of Snedecor and Cochran (1980) and Johnson and Leone (1977) are suggested; the books by Siegel (1956), and by Hollander and Wolfe (1973), provide useful presentations of many nonparametric methods.

Also included in the bibliography are references to several common packages of statistical computer programs. Perhaps the most versatile and universally available to the States is SAS. The SAS software is interfaced to STORET and may be used by STORET users. Documentation of how to use SAS within STORET is available through the EPA. Where possible, references are made in this document to the available procedures in SAS, for readers familiar with or having access to SAS.

Finally, it should be emphasized that the examples used in the following sections are hypothetical examples. Their purpose is to demonstrate the different statistical techniques and how to arrive at a test statistic necessary to perform a specific test. The reader should not be lead to believe that one year of monthly data, for example, is sufficient to apply a given test, because of the possibility of seasonal fluctuations.

IV. PARAMETRIC PROCEDURES

A. Methods to Deseasonalize Data

If the data exhibit a seasonal cycle--typically an annual cycle for monthly or quarterly data--then this seasonal effect must be removed or accounted for before testing for a long-term trend or a step trend. The method proposed here is simple and straightforward. Assume that the data are monthly values. To remove the possible seasonal effect, calculate the mean of the observations taken in different years, but during the same months. That is, calculate the mean of the measurements in January, then the mean for February, and so on for each of the 12 months.

After calculating the 12 monthly means, subtract the monthly mean from each observation taken during that month. These differences will then have any seasonal effects removed--the differences are thus deseasonalized data. If data were taken on a quarterly basis, one would calculate the four quarterly means, then subtract the mean for the first quarter from each of the first quarter observations, subtract the mean for the second quarter from each of the second quarter observations and so on. The resulting differences can be used subsequently for testing for a long-term trend or a step trend.

In the parametric framework, deseasonalizating can also be accomplished by using multiple regression with indicator variables for the months. Using multiple regression, as described in Subsection D below, this method of seasonal adjustment can be performed at the same time that a regression line is fitted to the data.

Many other approaches to deseasonalize data exist. If the seasonal pattern is regular it may be modeled with a sine or cosine function. Moving averages can be used, or differences (of order 12 for monthly data) can be used. Time series models may include rather complicated methods for deseasonalizing the data. However, the method described above should be adequate for the water quality data. It has the advantage of being easy to understand

and apply, and of providing natural estimates of the monthly effects via the monthly means.

B. Regression Analysis

A parametric procedure commonly used to test for trends is regression analysis. As stated earlier, however, water quality data often do not meet the underlying assumptions such as constant variance and normally distributed error terms, so regression analysis should be used with great caution. An analysis of the residuals (i.e., differences between the observed and the regression-predicted values of the water quality parameter) is therefore recommended.

With this proviso in mind, the following example will illustrate the method. Suppose an annual water quality index were available for a 7-year period.

Year:	1977	1978	1979	1980	1981	1982	1983
Year No.:	1	2	3	4	5	6	7
WQI:	46	52	42	44	39	45	40

We can use the years themselves in the numerical calculations which follow, but it is much easier and totally equivalent to just number them from 1 to 7 and use the numbers instead. Note that if a year is missing, then the other years should be numbered accordingly (i.e., the corresponding number will be missing also).

First, it is often a good idea to plot the data. Figure 6 shows these sample data as points on a plot of WQI versus year, along with two calculated points described subsequently.

The calculations with the data to determine the regression line, and to test it for statistical significance (i.e., whether the slope and intercept of the calculated line are statistically different from zero), can be done

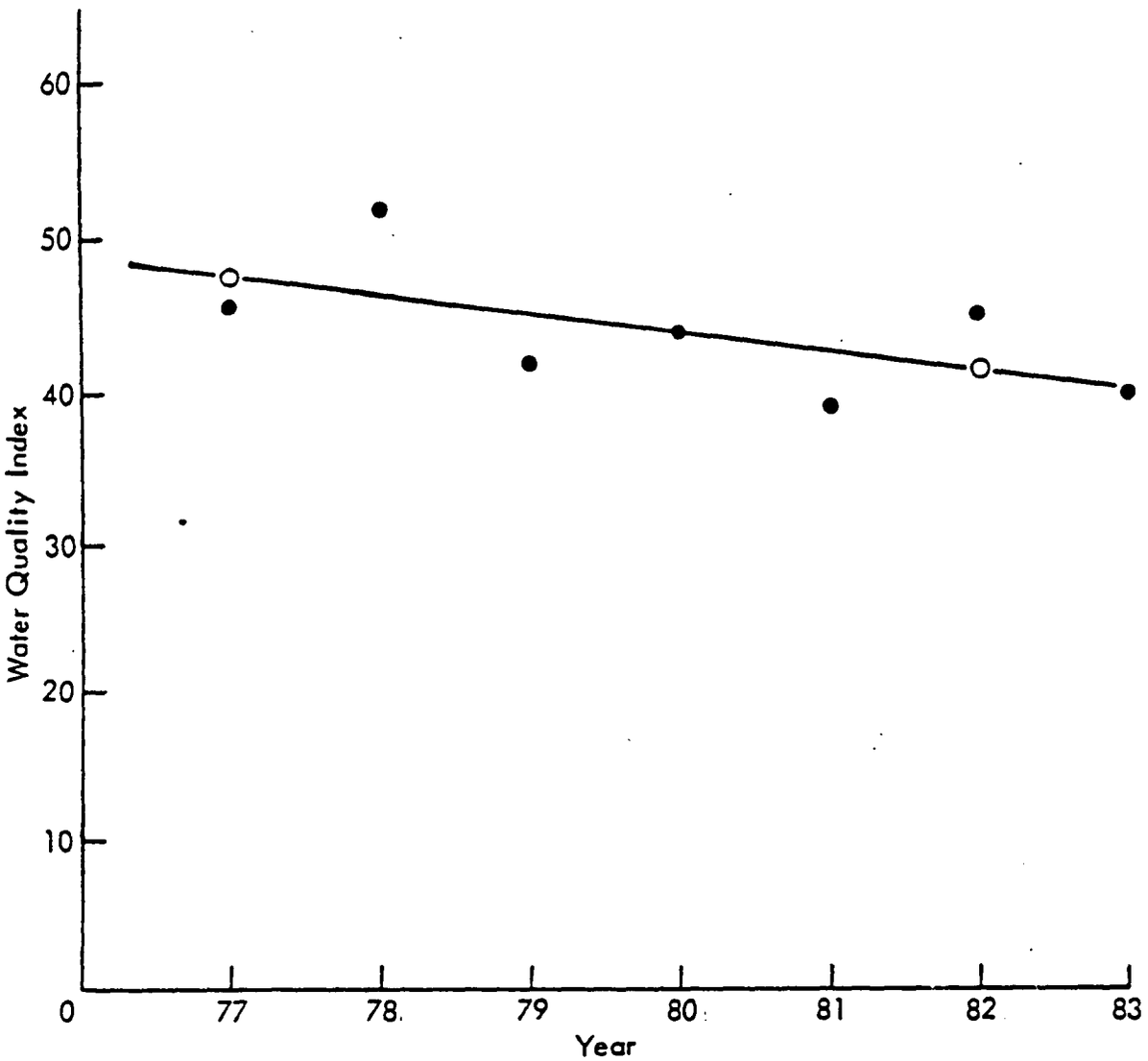


Figure 6 - Sample Linear Regression Line

manually. However, the formulas are fairly lengthy and the computations are quite tedious and therefore error-prone. These calculations are almost always done on a computer or hand calculator. Most scientific hand calculators, in fact, have the formulas built in, so we only have to enter the data, pairs of year and WQI, into the calculator and then read out the answers.

The equation for the regression line is:

$$Y = a + bX$$

where X is the year number (or year), Y is the WQI; a is called the intercept of the line, and b is called the slope. Upon entering all the X and Y values (the exact entry procedure depends upon the specific hand calculator being used), we read out the values of a and b.

Using the sample data, we get a = 49 and b = -1.25. Thus, the fitted regression line is:

$$Y = 49 - 1.25X \ .$$

The regression line can be drawn by plotting two arbitrary points from this regression equation and connecting them with a straight line. For example, choosing X = 1,

$$Y = 49 - 1.25(1) = 47.75$$

and for X = 6,

$$Y = 49 - 1.25(6) = 41.5 \ .$$

These points are also plotted in Figure 6 (as small o's). The straight line through these two points is the calculated regression line.

The fitted regression equation can be used to predict the value of the water quality index Y which corresponds to a given value of X. The difference between the observed value of Y and the predicted value of Y for a given X is called the residual at this point X. In our example, we have

Year No.:	1	2	3	4	5	6	7
Observed WQI:	46	52	42	44	39	45	40
Predicted WQI:	47.75	46.50	45.25	44.00	42.75	41.50	40.25
Residual:	-1.75	5.50	-3.25	0	-3.75	3.50	-0.25

The residual measures the discrepancy between the observed value and the value obtained from the regression line. If the points were perfectly lined up, each residual would be zero and we would have a perfect fit. Note that the residuals always add up to zero (rounding off aside). It is these residuals that should be analyzed to test whether they are normally distributed. The normal distribution assumption can be graphically checked by either using probability paper as explained in Section VI-C or by performing a test of fit using a computer program such as the UNIVARIATE procedure in SAS.

In this example, the WQI appears to have a decreasing trend. But, we must then ask, is the slope of the regression line statistically significant? In other words, is the decrease in the water quality index, by an amount of 1.25 per year, significantly different from zero? To test for this, compute the ratio of the slope, b, to its standard deviation and test using a t-table. This testing procedure is automatically done in SAS as shown in the appendix. When using a hand calculator, however, we use an equivalent test via r, the sample correlation coefficient. The Student's variable t with n-2 degrees of freedom is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} .$$

In the example data, r = -0.619 and n = 7, and we obtain t = 1.762. Using Table 2 below and reading at the intersection of the line headed by 5 (7-2 degrees of freedom) and the column headed 0.050 (two-sided significance level of 5%), we read a critical value of 2.571. Since t of 1.762 falls between -2.571 and +2.571, we conclude that the correlation coefficient, r, and consequently the slope, b, are not statistically different from zero, or that the apparent downward trend is not statistically significant.

More details on regression analysis are available in textbooks by Draper and Smith (1981), first chapter, and Chatterjee and Price (1977) who also provide extended discussions on the analysis of residuals.

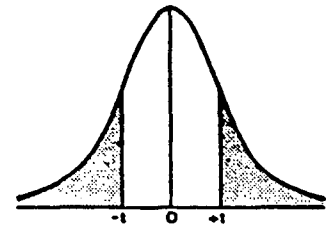
Regression analysis can be performed with SAS using one of several procedures with their appropriate options. One SAS procedure would be PROC REG which features an extensive output (see SAS User's Guide: Statistics, p. 40). An example with output is included in the appendix.

C. Student's T-Test

Student's t-test is probably the most widely used parametric test to compare two sets of data to determine whether the populations from which they come have means that differ significantly from each other. This test would commonly be used to determine if a change has occurred, as reflected in data obtained before and after some event such as the coming on-line of a sewage treatment plant. Student's t-test, being a parametric procedure, makes assumptions about the underlying distribution of the population from which the data are a sample. The basic assumption required to formally develop this test procedure is that the population be normally distributed; however, moderate departures from normality will not seriously affect the results. When using a t-test, one should assure that this basic assumption is not violated. Mathematical transformations of the data--e.g., log transform, exponential transform, etc.--can often be helpful in order to arrive at a normally distributed sample. In comparing two samples, an additional requirement besides normality is that the two distributions have equal variances.

TABLE 2

TWO-TAILED* SIGNIFICANCE LEVELS OF STUDENT'S T



Degrees of Freedom	Probability of a Larger Value, Sign Ignored								
	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657		
2	0.816	1.061	1.886	2.920	4.303	6.205	9.925	14.089	31.598
3	.765	0.978	1.638	2.353	3.182	4.176	5.841	7.453	12.941
4	.741	.941	1.533	2.132	2.776	3.495	4.604	5.598	8.610
→ 5	.727	.920	1.476	2.015	2.571	3.163	4.032	4.773	6.859
6	.718	.906	1.440	1.943	2.447	2.969	3.707	4.317	5.959
7	.711	.896	1.415	1.895	2.365	2.841	3.499	4.029	5.405
8	.706	.889	1.397	1.860	2.306	2.752	3.355	3.832	5.041
9	.703	.883	1.383	1.833	2.262	2.685	3.250	3.690	4.781
10	.700	.879	1.372	1.812	2.228	2.634	3.169	3.581	4.587
11	.697	.876	1.363	1.796	2.201	2.593	3.106	3.497	4.437
12	.695	.873	1.356	1.782	2.179	2.560	3.055	3.428	4.318
13	.694	.870	1.350	1.771	2.160	2.533	3.012	3.372	4.221
14	.692	.868	1.345	1.761	2.145	2.510	2.977	3.326	4.140
15	.691	.866	1.341	1.753	2.131	2.490	2.947	3.286	4.073
16	.690	.865	1.337	1.746	2.120	2.473	2.921	3.252	4.015
17	.689	.863	1.333	1.740	2.110	2.458	2.898	3.222	3.965
18	.688	.862	1.330	1.734	2.101	2.445	2.878	3.197	3.922
19	.688	.861	1.328	1.729	2.093	2.433	2.861	3.174	3.883
20	.687	.860	1.325	1.725	2.086	2.423	2.845	3.153	3.850
21	.686	.859	1.323	1.721	2.080	2.414	2.831	3.135	3.819
22	.686	.858	1.321	1.717	2.074	2.406	2.819	3.119	3.792
23	.685	.858	1.319	1.714	2.069	2.398	2.807	3.104	3.767
24	.685	.857	1.318	1.711	2.064	2.391	2.797	3.090	3.745
25	.684	.856	1.316	1.708	2.060	2.385	2.787	3.078	3.725
26	.684	.856	1.315	1.706	2.056	2.379	2.779	3.067	3.707
27	.684	.855	1.314	1.703	2.052	2.373	2.771	3.056	3.690
28	.683	.855	1.313	1.701	2.048	2.368	2.763	3.047	3.674
29	.683	.854	1.311	1.699	2.045	2.364	2.756	3.038	3.659
30	.683	.854	1.310	1.697	2.042	2.360	2.750	3.030	3.646
35	.682	.852	1.306	1.690	2.030	2.342	2.724	2.996	3.591
40	.681	.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551
45	.680	.850	1.301	1.680	2.014	2.319	2.690	2.952	3.520
50	.680	.849	1.299	1.676	2.008	2.310	2.678	2.937	3.496
55	.679	.849	1.297	1.673	2.004	2.304	2.669	2.925	3.476
60	.679	.848	1.296	1.671	2.000	2.299	2.660	2.915	3.460
70	.678	.847	1.294	1.667	1.994	2.290	2.648	2.899	3.435
80	.678	.847	1.293	1.665	1.989	2.284	2.638	2.887	3.416
90	.678	.846	1.291	1.662	1.986	2.279	2.631	2.878	3.402
100	.677	.846	1.290	1.661	1.982	2.276	2.625	2.871	3.390
120	.677	.845	1.289	1.658	1.980	2.270	2.617	2.860	3.373
∞	.6745	.8416	1.2816	1.6448	1.9600	2.2414	2.5758	2.8070	3.2905

*for a one-tailed significance level α, read in column 2α.

df = n-1

Source: Snedecor and Cochran (1980) p. 469

This means that the distributions of the two populations are identical in shape, although they may differ in location (in other words, only their means may differ).

Procedure

The following is a series of 18 concentration measurements for total chromium, 8 before and 10 after implementation of a pollution control measure.

"Before" Concentrations (μg/ℓ)	"After" Concentrations (μg/ℓ)
99	59
111	99
74	82
123	51
71	48
75	39
59	42
85	42
	47
	50

First it is necessary to calculate the mean and standard deviation of each of the data sets. The mean (or average) is just the sum of all the values divided by the number of values, n. Thus, for the "before" data, the mean, m_B, is:

$$m_B = (99 + 111 + \ldots + 85)/8 = 87.1 \text{ μg/ℓ}$$

The standard deviation involves using the sum of the squares of each of the values. Specifically,:

$$S_B = \sqrt{\frac{n_B\,(99^2 + 111^2 + \ldots + 85^2) - (99 + 111 + \ldots + 85)^2}{n_B\,(n_B - 1)}}$$

where n_B is the number of "before" data points (8). Carrying out all the calculations gives:

$$S_B = \sqrt{481.8} = 22.0$$

Thus, we can compute:

"Before"	$m_B = 87.1$ µg/ℓ;	$S_B = 22.0$ µg/ℓ;	$S_B^2 = 481.8$;	$n_B = 8$
"After"	$m_A = 55.9$ µg/ℓ;	$S_A = 19.5$ µg/ℓ;	$S_A^2 = 380.1$;	$n_A = 10.$

The t-test for a step trend in this data set is simply a method of estimating whether the means m_B and m_A of the two partitions ("before" and "after") of the data set differ significantly at a chosen level of significance α. The test statistic t is:

$$t = \frac{m_B - m_A}{S_p \sqrt{1/n_B + 1/n_A}}$$

where S_p, the pooled standard deviation, is computed as:

$$S_p = \sqrt{\frac{n_B - 1 \;\; S_B^2 + n_A - 1 \;\; S_A^2}{n_B + n_A - 2}}$$

assuming S_B^2 and S_A^2 are not statistically different. To test whether the "before" concentrations are on the average higher than the "after" concentrations (i.e., one-sided test) at the α-level of significance, we compare the computed t with tabulated values of the Student's t distribution at probability level $(1-\alpha)$ with $(n_B + n_A - 2)$ degrees of freedom.

In the example,

$$S_p = \sqrt{\frac{7 \cdot 481.8 + 9 \cdot 380.1}{16}} = \sqrt{424.6} = 20.6 \ ,$$

thus:
$$t = \frac{87.1 - 55.9}{20.6\sqrt{1/8 + 1/10}} = \frac{31.2}{9.8} = 3.18$$

The critical t at $\alpha = 0.05$ (5%) with (8 + 10 - 2) = 16 degrees of freedom for a one-sided test is $t_c = 1.746$ (see Table 2 above). Since the computed t of 3.18 is larger, one concludes that, indeed, the "before" concentrations were higher on the average than the "after" concentrations, or that there is a significant decreasing step trend in the data at the 95% confidence level when the new treatment facility went into operation.

In the above example, it was assumed (without proof) that the variances, S_B^2 and S_A^2, were not statistically different. This assumption can (and should) be checked by using an F-test. Calculate:

$$F = \frac{S_B^2}{S_A^2} = \frac{481.8}{380.1} = 1.27 .$$

Note that due to tabulation restrictions, F is always computed as the larger variance over the smaller. That is, if S_A were greater than S_B, then $F = S_A^2/S_B^2$.

If the population variances are the same, one would expect F = 1. To test whether the calculated F is statistically greater than 1, an F table is used. Using Table 3 below, which is applicable at the $\alpha = 0.05$ level of significance, look up the critical F value in the column headed by 7 ($= n_B-1$) and the row labeled 9 ($=n_A-1$). The critical F value in this example is 3.29. If the computed F were greater than 3.29, then we would say with 95% confidence that the two variances were different, and so the Student's t-test could not be used. In our case F is less than the critical value, we therefore cannot confidently say that they are different, and so we accept the assumption of equal variances in the "before" and "after" groups.

In the case of unequal variances, when the Student's t-test cannot be used, only an approximate t-test such as the Behrens-Fisher test (Snedecor and Cochran (1980) p. 97) can be used to compare the two means.

TABLE 3

THE 5% (ROMAN TYPE) AND 1% (BOLDFACE TYPE) POINTS FOR THE DISTRIBUTION OF F

n_2	n_1 df in Numerator																							n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
1	161 **4,052**	200 **4,999**	216 **5,403**	225 **5,625**	230 **5,764**	234 **5,859**	237 **5,928**	239 **5,981**	241 **6,022**	242 **6,056**	243 **6,082**	244 **6,106**	245 **6,142**	246 **6,169**	248 **6,208**	249 **6,234**	250 **6,261**	251 **6,286**	252 **6,302**	253 **6,323**	253 **6,334**	254 **6,352**	254 **6,361**	254 **6,366**	1
2	18.51 **98.49**	19.00 **99.00**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.36**	19.37 **99.37**	19.38 **99.39**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**	19.42 **99.43**	19.43 **99.44**	19.44 **99.45**	19.45 **99.46**	19.46 **99.47**	19.47 **99.48**	19.47 **99.48**	19.48 **99.49**	19.49 **99.49**	19.49 **99.49**	19.50 **99.50**	19.50 **99.50**	2
3	10.13 **34.12**	9.55 **30.82**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**	8.71 **26.92**	8.69 **26.83**	8.66 **26.69**	8.64 **26.60**	8.62 **26.50**	8.60 **26.41**	8.58 **26.35**	8.57 **26.27**	8.56 **26.23**	8.54 **26.18**	8.54 **26.14**	8.53 **26.12**	3
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**	5.87 **14.24**	5.84 **14.15**	5.80 **14.02**	5.77 **13.93**	5.74 **13.83**	5.71 **13.74**	5.70 **13.69**	5.68 **13.61**	5.66 **13.57**	5.65 **13.52**	5.64 **13.48**	5.63 **13.46**	4
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.29**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**	4.64 **9.77**	4.60 **9.68**	4.56 **9.55**	4.53 **9.47**	4.50 **9.38**	4.46 **9.29**	4.44 **9.24**	4.42 **9.17**	4.40 **9.13**	4.38 **9.07**	4.37 **9.04**	4.36 **9.02**	5
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**	3.96 **7.60**	3.92 **7.52**	3.87 **7.39**	3.84 **7.31**	3.81 **7.23**	3.77 **7.14**	3.75 **7.09**	3.72 **7.02**	3.71 **6.99**	3.69 **6.94**	3.68 **6.90**	3.67 **6.88**	6
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**	3.52 **6.35**	3.49 **6.27**	3.44 **6.15**	3.41 **6.07**	3.38 **5.98**	3.34 **5.90**	3.32 **5.85**	3.29 **5.78**	3.28 **5.75**	3.25 **5.70**	3.24 **5.67**	3.23 **5.65**	7
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**	3.23 **5.56**	3.20 **5.48**	3.15 **5.36**	3.12 **5.28**	3.08 **5.20**	3.05 **5.11**	3.03 **5.06**	3.00 **5.00**	2.98 **4.96**	2.96 **4.91**	2.94 **4.88**	2.93 **4.86**	8
→ 9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**	3.02 **5.00**	2.98 **4.92**	2.93 **4.80**	2.90 **4.73**	2.86 **4.64**	2.82 **4.56**	2.80 **4.51**	2.77 **4.45**	2.76 **4.41**	2.73 **4.36**	2.72 **4.33**	2.71 **4.31**	9
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**	2.86 **4.60**	2.82 **4.52**	2.77 **4.41**	2.74 **4.33**	2.70 **4.25**	2.67 **4.17**	2.64 **4.12**	2.61 **4.05**	2.59 **4.01**	2.56 **3.96**	2.55 **3.93**	2.54 **3.91**	10
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	3.01 **4.88**	2.95 **4.74**	2.90 **4.63**	2.86 **4.54**	2.82 **4.46**	2.79 **4.40**	2.74 **4.29**	2.70 **4.21**	2.65 **4.10**	2.61 **4.02**	2.57 **3.94**	2.53 **3.86**	2.50 **3.80**	2.47 **3.74**	2.45 **3.70**	2.42 **3.66**	2.41 **3.62**	2.40 **3.60**	11
12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.92 **4.65**	2.85 **4.50**	2.80 **4.39**	2.76 **4.30**	2.72 **4.22**	2.69 **4.16**	2.64 **4.05**	2.60 **3.98**	2.54 **3.86**	2.50 **3.78**	2.46 **3.70**	2.42 **3.61**	2.40 **3.56**	2.36 **3.49**	2.35 **3.46**	2.32 **3.41**	2.31 **3.38**	2.30 **3.36**	12
13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.84 **4.44**	2.77 **4.30**	2.72 **4.19**	2.67 **4.10**	2.63 **4.02**	2.60 **3.96**	2.55 **3.85**	2.51 **3.78**	2.46 **3.67**	2.42 **3.59**	2.38 **3.51**	2.34 **3.42**	2.32 **3.37**	2.28 **3.30**	2.26 **3.27**	2.24 **3.21**	2.22 **3.18**	2.21 **3.16**	13

n_1: degrees of freedom in Numerator of F

n_2: degrees of freedom in Denominator of f

Source: Snedecor and Cochran (1980) p. 480

A final note: many hand calculators will automatically compute means and standard deviations, making the above calculations quite easy. Also, the t-test can be performed using SAS. The procedure PROC TTEST may be used (see SAS User's Guide: Statistics, p. 217). An example output is shown in the appendix.

D. Trend and Change

Over a long period of time, a situation could arise where both a long-term trend and a step trend due to the implementation of a new treatment facility might be present in the series of data. In such a case, the two methods described above (regression analysis and t-test procedure), would be combined to test for both types of trends. Practically, this is done by performing a multiple regression analysis where the dependent variable is the water quality parameter of interest, one independent variable is the time (e.g., month) and the second independent variable is a "dummy" variable taking the value 0 before the known change and 1 after the change. In addition, if a seasonal effect is suspected (this might be detected when plotting the data), then the data need to be deseasonalized by introducing 12 indicator variables. The first one would take the value 1 for January and 0 otherwise; the second would be 1 for February and 0 otherwise, etc. The regression equation can be written as:

$$Y = a_1X_1 + a_2X_2 + \ldots + a_{12}X_{12} + bC + cT$$

where a_i is the effect of month i, i = 1,...,12
b is the change effect, and
c is the slope for the long-term trend.

Algebraically, the above equation is equivalent to

$$Y - a_1X_1 - a_2X_2 - \ldots - a_{12}X_{12} = bC + cT.$$

The left-hand side of this equation is just the deseasonalized data, while the right-hand side is the contribution due to change and trend. The next

step is to estimate all 14 parameters (a_1, ..., a_{12}, b and c) via least squares method and to test whether b and c (the coefficients for change and trend, respectively) are significantly different from zero.

A complete example, based on the data plotted in Figure 1 of Section III, is shown in the appendix using the multiple regression procedure of SAS.

E. Time Series Analysis

Time series analysis is a set of parametric procedures that can be applied to a series of ordered observations of a quantitative measure, such as a water quality indicator, taken at regular points in time. Although not essential, it is common for the points to be equally spaced in time, for example, monthly. The objective in time series analysis is to determine from the set of data the pattern of change over time (e.g., time trends, seasonality, cyclical variations, etc.). The various measured patterns are extracted one by one until the remaining variation in the data is purely random. When this has been done, all the meaningful information contained in the original data has been "captured," and the random component that remains is, by definition, worthless for forecasting. In general, the various trends and patterns can then be used to forecast probable future behavior of the series, and confidence intervals can be computed for future projections.

Various analytical methods have been developed to decompose a time series into trend, seasonal, change, and irregular components. These methods are fairly complex, nearly always require special computer programs, and thus are beyond the scope of these guidelines. The interested reader is referred to standard textbooks on time series such as Box and Jenkins (1970), Kendall and Stuart (1966), and Glass et al. (1975) or to the publications by Box and Tiao (1975) and Schlicht (1981).

Time series analysis is not suitable for use with many water quality data bases because of missing data, values reported as below detection limits, and changing laboratory techniques (see van Belle and Hughes (1982)). Also,

time series methods generally require several years of monthly observations. For these reasons, the methods suggested here are generally more practical for analyzing water quality data.

V. DISTRIBUTION-FREE METHODS

A. Runs Tests for Randomness

Given a series of measurements of a water quality parameter or indicator derived therefrom, the question might be asked if the data vary in a random manner or if they indicate a long-term trend or a seasonal or other periodic fluctuation. An easy test to detect nonrandom components is the runs test illustrated in the following simple example. For a given constituent, an average monthly water quality index (WQI) has been computed and categorized as either bad (B) or good (G). The figures are:

Month:	1	2	3	4	5	6	7	8	9	10	11	12
WQI:	G	G	G /	B	B	B	B	B /	G	G	G	G

In statistical terms, the above question can be formulated as a null hypothesis: the G's and B's occur in random order; versus an alternative: the order of the G's and B's deviates from randomness.

Each cluster of like observations is called a run. Thus, there are 3 runs in the series of data above. Let n_1=5 be the number of B's, and n_2 = 7 the number of G's (n_1 denotes the smaller of the two numbers). If there are too many or too few runs, then the assumption that the sample is random is probably not correct. If very few runs occur, a time trend or some bunching due to lack of independence is suggested. If a great many runs occur, systematic short-period cyclical fluctuations seem to be influencing the WQI data. Tables have been developed for n_1 and n_2 up to 20 showing probability levels for the runs test, and can be found in Langley (1971). An example showing 5% probability levels is included in Table 4. For n_1 = 5 and n_2 = 7 (at arrow) we use the table as follows: reject the null hypothesis that the sample is a random ordering of G's and B's if the observed number of runs is equal to or less than the smaller number in the table (i.e., 3), or is equal to or greater than the larger number in the table (i.e., 11). Since we have a series with three runs, we conclude that the fluctuations are not random;

TABLE 4

r TABLES SHOWING 5% LEVELS FOR RUNS TEST

n_1	n_2	No. of runs	
2	2-11	—	—
2	12-20	2	—
3	3-5	—	—
3	6-14	2	—
3	15-20	3	—
4	4	—	—
4	5-6	2	9
4	7	2	—
4	8-15	3	—
4	16-20	4	—
5	5	2	10
5	6	3	10
5	7-8	3	11
5	9-10	3	—
5	11-17	4	—
5	18-20	5	—
6	6	3	11
6	7-8	3	12
6	9-12	4	13
6	13-18	5	—
6	19-20	6	—
7	7	3	13
7	8	4	13
7	9	4	14
7	10-12	5	14
7	13-14	5	15
7	15	6	15
7	16-20	6	—
8	8	4	14
8	9	5	14
8	10-11	5	15
8	12-15	6	16
8	16	6	17
8	17-20	7	17
9	9	5	15
9	10	5	16
9	11-12	6	16
9	13	6	17
9	14	7	17
9	15-17	7	18
9	18-20	8	18
10	10	6	16
10	11	6	17
10	12	7	17
10	13-15	7	18

n_1	n_2	No. of runs	
10	16-18	8	19
10	19	8	20
10	20	9	20
11	11	7	17
11	12	7	18
11	13	7	19
11	14-15	8	19
11	16	8	20
11	17-18	9	20
11	19-20	9	21
12	12	7	19
12	13	8	19
12	14-15	8	20
12	16-18	9	21
12	19-20	10	22
13	13	8	20
13	14	9	20
13	15-16	9	21
13	17-18	10	22
13	19-20	10	23
14	14	9	21
14	15	9	22
14	16	10	22
14	17-18	10	23
14	19	11	23
14	20	11	24
15	15	10	22
15	16	10	23
15	17	11	23
15	18-19	11	24
15	20	12	25
16	16	11	23
16	17	11	24
16	18	11	25
16	19-20	12	25
17	17	11	25
17	18	12	25
17	19	12	26
17	20	13	26
18	18	12	26
18	19	13	26
18	20	13	27
19	19-20	13	27
20	20	14	28

Source: Langley (1971) p. 325

that is, there is only a small probability of obtaining three or fewer runs (less than once in 20 times) if the sample was actually random. Should the observed number of runs fall between the two values given in the table, then we can accept the sample as being random. If we come across a dash (-) in the table, it means that a 5% probability level cannot be reached in the particular circumstances, regardless of the number of runs.

If the runs test is used for data sets where n_2 is greater than 20, then rather than using these special types of tables, a different approach is used where a Z-statistic is computed. The formula is:

$$Z = \frac{\left| r - \frac{2n_1n_2}{N} + 1 \right|}{\sqrt{\left(\frac{2n_1n_2}{N}\right)\left(\frac{2n_1n_2 - N}{N^2 - N}\right)}}$$

where N is simply $(n_1 + n_2)$, and r is the number of runs (3 in the above example).

The notation, $|\ |$, means use the absolute value of what is within--that is, make the resulting number positive. For example, if $n_1 = 6$, $n_2 = 24$, and there are 10 runs, then:

$$\left| r - \frac{2n_1n_2}{N} + 1 \right| = \left| 10 - \frac{2(6)(24)}{30} + 1 \right| = \left| 10 - 10.6 \right| = +0.6$$

The denominator in Z equals: $\sqrt{\left(\frac{2(6)(24)}{30}\right)\left(\frac{2(6)(24) - 30}{30^2 - 30}\right)} = 2.85.$

Thus, $Z = \frac{0.6}{2.85} = 0.21.$

Reference to Table 1 (page 19) shows that to reject the hypothesis at a two-sided significance level of 0.05 requires a Z-value of 1.96 or greater. Since we obtained Z = 0.21, which is less than the critical value of 1.96, we cannot reject the null hypothesis of the sample being random.

Note: It should be emphasized that this latter approach, which involves computing a Z-statistic and comparing it with the tabulated standard normal deviates in Table 1, is only to be used with larger data sets, where n_2 exceeds 20.

The above example was based on two kinds of observations (good, bad) only. However, the runs test applies to any kind of data that can be categorized into two groups. If the data are quantitative measurements, first find the median. Then assign a plus sign whenever the observation is above the median, and a minus sign whenever the observation is below the median. Then proceed as above, using the plus/minus categories in the same way as the good/bad categories were used.

If a trend is the alternative to randomness that is of particular importance, then a runs test appropriate to this alternative can be constructed as follows. Count as a plus each observation that exceeds the preceding one; count as a minus each observation that is less than the preceding one. Then use these plus/minus categories as before. By changing the definition of the two categories one can arrive at different runs tests for different alternatives to randomness.

Many types of data, however, have more than two categories. For example, a water quality index could be partitioned into "very good," "good," "fair," "poor," and "very poor" categories. The above mentioned runs test for two categories has been generalized to cases with any number of categories. The appropriate test statistics are derived Z-statistics and can therefore be tested by using the tabulated standard normal deviates. Details are given in Wallis and Roberts (1956), Chapter 18.

The runs test can detect a wide variety of departures from randomness. However, if a specific type of departure is important, a test designed to detect that type of departure will be more likely to detect it. With water quality data taken over time there are three particular types of departures from randomness that are of interest. One is a seasonal effect related to

different amounts of precipitation or weather changes over the year. A more interesting effect is a long-term trend in which the water quality level is gradually improving (or worsening) over time. Finally, a third is a sudden change in water quality associated with a discrete event--opening of a new factory or installation of a new treatment plant. Of these three, seasonality is generally a nuisance. That is, one is not specifically interested in it but rather must adjust for it before the other effects (long-term trend or step trend) can be tested for. Specific tests for these types of trend are presented in the following subsections.

B. Kendall's Tau Test

Kendall's tau is a rank correlation coefficient. A distribution-free test based on Kendall's statistic is commonly used to compare one set of numerical data with another to see if they tend to "track" together. In water quality work, a series of readings of some parameter can be tested against time (e.g., the series of months over which the parameter was obtained) to see if it has any generally increasing or decreasing tendency. This tendency does not need to be linear (i.e., straight line). To illustrate the method, assume one has the following 12 monthly average water quality indices (WQIs) at a given station:

Month:	1	2	3	4	5	6	7	8	9	10	11	12
WQI:	21	3	5	8	21	48	37	39	26	16	35	7

Statistically speaking, one may wish to test the null hypothesis that there is no trend in the data, i.e., months and WQI values are unrelated, against the alternative that there is a trend in the data, i.e., the two variables are related.

The first step is to rank the months and the WQIs in order from lowest to highest. Since the months are already in order, it is only necessary to rank the WQIs. The smallest is 3 (in month 2), so it is ranked number 1. Similarly, the second WQI in rank is 5, then 7, etc. Note that

the value 21 appears twice, so the WQIs for months 1 and 5 are "tied." This is a common situation with this type of data, and is resolved by averaging the ranks. In this case, the two WQIs in question share ranks 6 and 7, so each is given the average value, (6+7)/2 = 6.5. The final rankings are tabulated below.

Month:	1	2	3	4	5	6	7	8	9	10	11	12
WQI:	21	3	5	8	21	48	37	39	26	16	35	7
Rank:	6.5	1	2	4	6.5	12	10	11	8	5	9	3
k+:	-	0	1	2	3	5	5	6	5	3	7	2
k-:	-	1	1	1	0	0	1	1	3	6	3	9

The test involves determining the extent to which the set of WQI values is ordered in the same way as the months. The following explains the procedure.

Take each WQI rank and count how many of the ranks to the left of it are smaller; this gives the k+ line. Then sum up the 11 k+ values; this yields K+, the number of concordant pairs (i.e., pairs ordered in the same way as the months). Then repeat, but count how many of the values to the left of each WQI rank are greater; this gives 11 k- values. The sum of these, K-, is the number of discordant pairs. The tie at (21,21) is disregarded in the counts of K+ and K-. Kendall's tau is then computed as:

$$\text{tau} = \frac{(K+) - (K-)}{n(n-1)/2} \quad .$$

If there were no ties and concordant pairs only, then K+ = n(n-1)/2, K- = 0, and tau = 1; if there were discordant pairs only and no ties, then K+ = 0, K- = n(n-1)/2, and tau = -1. If K+ = K-, then tau = 0.

In our example,

$$K+ = (0 + 1 + 2 \ .... + 3 + 7 + 2) = 39,$$

$$K- = (1 + 1 + ... + 3 + 9) = 26, \text{ and}$$

$$\text{tau} = \frac{39-26}{(12)(11)/2} = 0.197 \ .$$

For small sample sizes, n, the significance of tau is tested by means of tabulations of values of Kendall's K = (K+)-(K-), rather than of tau itself. There is a series of tables for sample sizes ranging from n = 4 to 40 (Hollander and Wolfe, 1973). Table 5 shows a sample of these tabulations. For larger values of n a normal approximation, discussed subsequently, is used.

Our example yields a K of (39-26) = 13. In Table 5, under Column n = 12 (number of observations), we read 0.230 at x = 12 and 0.190 at x = 14. Thus, the one-tailed probability associated with K = 13 is about 0.21 (midway between 0.230 and 0.190); the two-tailed significance level associated with K = 13 is thus 2 · 0.21 = 0.42. If we had initially chosen a desired level of significance of 5% (0.05), we would conclude that there is no significant trend in the data at this level since 0.42 is not small enough (it is greater than 0.05).

We could also use Table 5 to determine the 5% critical value for n = 12. In Column n = 12 we read down to find 0.022 (closest probability to 0.05/2 = 0.025), then read across to x = 30. The probability of 0.031 in the same column yields x = 28. The critical K associated with an approximate two-sided 5% significant level is therefore between 28 and 30, so one could use 29, (or -29 if K were negative). Since K of 13 lies between -29 and +29, we come to the same conclusion as above, i.e., since K is not big enough, nor small enough, there is no significant trend. Note that if there are no ties, K can take on only even values. This is why there is no entry for K=29 in the table.

TABLE 5

UPPER TAIL PROBABILITIES FOR THE NULL DISTRIBUTION OF KENDALL'S K STATISTIC (Subtable)

	n →								
x	4	5	8	9	12	13	16	17	20
0	.625	.592	.548	.540	.527	.524	.518	.516	.513
2	.375	.408	.452	.460	.473	.476	.482	.484	.487
4	.167	.242	.360	.381	.420	.429	.447	.452	.462
6	.042	.117	.274	.306	.369	.383	.412	.420	.436
8		.042	.199	.238	.319	.338	.378	.388	.411
10		.008	.138	.179	.273	.295	.345	.358	.387
→ 12			.089	.130	(.230)	.255	.313	.328	.362
→ 14			.054	.090	(.190)	.218	.282	.299	.339
16			.031	.060	.155	.184	.253	.271	.315
18			.016	.038	.125	.153	.225	.245	.293
20			.007	.022	.098	.126	.199	.220	.271
22			.002	.012	.076	.102	.175	.196	.250
24			.001	.006	.058	.082	.153	.174	.230
26			.000	.003	.043	.064	.133	.154	.211
28				.001	.031	.050	.114	.135	.193
30				.000	.022	.038	.097	.118	.176
32					.016	.029	.083	.102	.159
34					.010	.021	.070	.088	.144
36					.007	.015	.058	.076	.130
38					.004	.011	.048	.064	.117
40					.003	.007	.039	.054	.104
42					.002	.005	.032	.046	.093
44					.001	.003	.026	.038	.082
46					.000	.002	.021	.032	.073
48						.001	.016	.026	.064
50						.001	.013	.021	.056
52						.000	.010	.017	.049
54							.008	.014	.043
56							.006	.011	.037
58							.004	.009	.032
60							.003	.007	.027
62							.002	.005	.023
64							.002	.004	.020
66							.001	.003	.017
68							.001	.002	.014
70							.001	.002	.012

Source: Hollander and Wolfe (1973) p. 384-393

For large sample sizes, a normal distribution approximation to the distributions of K or tau may be considered. Under the null hypothesis, the expected value of tau is 0 and the variance of tau is 2(2n+5)/9n(n-1). Thus, the ratio:

$$Z = \frac{\text{tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}}$$

is approximately standard normally distributed and Table 1 (page 19) is used for testing.

Our example, based on a tau of 0.197, would yield a value

$$Z = \frac{0.197}{\sqrt{\frac{2(24+5)}{(9)(12)(11)}}} = 0.89 \ .$$

In Table 1 we read a one-tailed probability of 0.1867 at Z = 0.89 from which a two-tailed probability of twice 0.1867, or approximately 0.37, follows. Again, the computed value of Z of 0.89 is not significant at the 5% level and the same conclusion is reached. Note that from Table 5 we obtained the exact probability of 0.42; the large sample size approximation yielded a probability of 0.37. The discrepancy between the two probabilities arises because the large sample size approximation was used for a sample size of only 12.

In case of ties, the value of tau as defined above is not affected. However, the variance of tau when using the large sample size approximation needs a correction for ties. The correction is lengthy in form and has little effect when only a few ties are present (see Hollander and Wolfe, 1973, p. 187).

This test can be performed using the SAS procedure, PROC CORR, with the KENDALL option (SAS User's Guide: Basics, p. 501). An example output with the corresponding SAS statements is presented in the appendix.

Once a trend is found to be significant, the next step is to estimate its magnitude. A common measure of the magnitude of the trend is the slope of a straight line fitted to the data. To find the distribution-free estimate of the slope, calculate the slopes for all possible pairs of observations. That is, calculate

$$S_{ij} = (Y_j - Y_i)/(X_j - X_i)$$

for i=1,2, ..., n-1 and j=i+1, i+2, ..., n. There are N=n(n-1)/2 such distinct pairs. For the example of WQI values used here, N=12(11)/2=66.

Then the N slopes are arranged in ascending order, and the middle value (the median) is the best estimate of the slope. In this case, the ordered series of slopes becomes -28, -18, -13, ..., 19, 20, 27. The median is the average of the 33rd and 34th values in the series, or 1.49. Thus, we would conclude that the WQI is, on the average, increasing by 1.49 units per month. However, it must be remembered that we found no significant trend, so that the value 1.49 is not significantly different from zero. Indeed, more advanced procedures also allow one to calculate confidence bounds from these ordered values; in this case the approximate 95% confidence interval for the slope is (-4.33, +3.85), a wide interval that includes both negative and positive possibilities. The interested reader may find further details in Hollander and Wolfe (1973). SAS does not include routines for estimating the magnitude and confidence bounds of a trend using Kendall's tau.

C. The Wilcoxon Rank Sum Test (Step Trend)

This is a distribution-free test most appropriate for testing for a so-called step trend. A step trend might be evident when some major event such as placing a new treatment facility into operation occurred during the data collection period.

Procedure

The general procedure involves comparing the rankings of one set of values (e.g., readings obtained before a major event) with the rankings of another set (e.g., readings obtained after a major event). There need not be an equal number of values in each set. Because of the way the tables for W, the Wilcoxon test statistic, are arranged, the set with the fewer number of values, (sample size n), is always compared to the larger set (sample size m), and not vice versa. Thus n is always the smaller of the two sample sizes.

Consider again the example used to demonstrate Student's t test, namely the series of 18 concentration measurements for total chromium, 8 before (denoted by X_i) and 10 after (denoted by Y_j) implementation of a pollution control measure.

"Before" Concentrations (µg/ℓ)	Ranks	"After" Concentrations (µg/ℓ)	Ranks
99	15.5	59	8.5
11	17	99	15.5
74	11	82	13
23	18	51	7
71	10	48	5
75	12	39	1
59	8.5	42	2.5
85	14	42	2.5
		47	4
		50	6

Let us first consider the following test situation: because we do not expect (or are not interested in) a worsening of the water quality due to the new, improved facility, we will test the null hypothesis that the "before" and "after" concentrations are equally high or low (i.e., there is no change) against the one-sided alternative that the "before" concentrations are higher than the "after" concentrations (i.e., there is an improvement in the water quality).

The first step, as with many other distribution-free procedures, is to determine the ranks of the concentration values. The observations are ordered as a single set of data (i.e., disregard "before" and "after"); if ties are present, use average ranks. Wilcoxon's test statistic is simply the sum of the ranks of the values in the group of smaller size (here the before group with n = 8):

$$W = (15.5 + 17 + ... + 14) = 106.$$

For a one-sided test at the α = 5% level of significance, we reject the null hypothesis if W is greater than or equal to the critical value, W_c, associated with m, n, and α, and we accept the null hypothesis otherwise. Table 6 gives the one-sided levels of significance for n = 8 and several values of m. Note that we might not always find the exact α that we have chosen, because the significance levels are discrete and so are only tabulated for integer values of m and n. In our example, for α = 0.051 (approximately 5%) we read that x = 95 (= W_c) for n = 8 and m = 10 (see arrows). Since W of 106 is greater than 95, we reject the null hypothesis of no change and conclude that the concentrations have significantly decreased after implementation of the improvement measure.

Also, Table 6 shows for n = 8 and m = 10 the probability of obtaining a value of W greater than or equal to 106 to be 0.003. This means that the probability of obtaining a W of 106 or more under the null hypothesis is 0.003, which is small enough to reject the hypothesis in favor of the alternative.

Next let us consider another testing situation. Suppose a new industrial discharge began, and the same "before" and "after" data as above were obtained. Here, we will test the null hypothesis of no change against the alternative that the water quality has degraded, i.e., the "before" concentrations are lower on the average than the "after" concentrations. (For these data, we already know this is not true. However, we will go through the analysis process anyway to illustrate the procedure.)

TABLE 6

UPPER TAIL PROBABILITIES FOR
WILCOXON'S RANK SUM W STATISTIC
(Subtable)

$n = 8$			
x	$m = 8$	$m = 9$	$m = 10$
68	.520		
69	.480		
70	.439		
71	.399		
72	.360	.519	
73	.323	.481	
74	.287	.444	
75	.253	.407	
76	.221	.371	.517
77	.191	.336	.483
78	.164	.303	.448
79	.139	.271	.414
80	.117	.240	.381
81	.097	.212	.348
82	.080	.185	.317
83	.065	.161	.286
84	.052	.138	.257
85	.041	.118	.230
86	.032	.100	.204
87	.025	.084	.180
88	.019	.069	.158
89	.014	.057	.137
90	.010	.046	.118

$n = 8$			
x	$m = 8$	$m = 9$	$m = 10$
91	.007	.037	.102
92	.005	.030	.086
93	.003	.023	.073
94	.002	.018	.061
95	.001	.014	.051
96	.001	.010	.042
97	.001	.008	.034
98	.000	.006	.027
99	.000	.004	.022
100	.000	.003	.017
101		.002	.013
102		.001	.010
103		.001	.008
104		.000	.006
105		.000	.004
106		.000	.003
107		.000	.002
108		.000	.002
109			.001
110			.001
111			.000
112			.000
113			.000
114			.000
115			.000
116			.000

Source: Hollander and Wolfe (1973) p. 272-282

The computation of W is unchanged (i.e., W = 106). However, we will reject the null hypothesis at the α level of significance whenever W is less than or equal to the critical value of $n(m+n+1) - W_c$, where W_c is defined as above, and accept the null hypothesis otherwise. Thus, with an α of 0.051, m = 10, and n = 8, $n(m+n+1) - W_c = (8)(19) - 95 = 57$, and we cannot reject the hypothesis in favor of a degradation since 106 is not less than the critical value of 57.

For a two-sided test of the null hypothesis of "no change" against the alternative of "a change at the 5% level of significance," we compare the above W of 106 with the following two critical values: $n(m+n+1) - W(\alpha_1, m, n)$ and $W(\alpha_2, m, n)$, where $\alpha_1 + \alpha_2 = \alpha$. Most often, we cannot perform the test at the exact α level, so we have to choose α_1 and α_2 from the table as close to $\alpha/2$ as possible. In our example, W(0.027, 10, 8) = 98 and W(0.022, 10, 8) = 99. Thus for $\alpha = 0.022 + 0.027 = 0.049$, the two critical values would be either (8)(19) - 98 = 54 and 99, or (8)(19) - 99 = 53 and 98. Since the computed W of 106 lies outside the interval 54-99 or 53-98, we reject the null hypothesis in favor of the alternative of a significant change.

The upper tail probabilities associated with smaller x-values (see Table 6) are not tabulated and are generally not of interest because the corresponding α-values are greater than 0.5. However, they can be calculated. If the probability that W is greater than or equal to x is $P(W \geq x)$, which is not tabulated because it is greater than 0.500, then calculate:

$$P(W \geq x) = 1 - P[W \geq (n(m+n+1) - x + 1)]$$

Example: n = 8, m = 10, x = 75:

$$P(W \geq 75) = 1 - P[W \geq (8(19) - 75 + 1)]$$

$$= 1 - P(W \geq 78)$$

$$= 1 - 0.448$$

$$= 0.552$$

Note: Wilcoxon's W statistic is tabulated in Hollander and Wolfe (1973), pp. 272-282 for values of m up to 20. Table 6 is just a sample of these tabulations. For larger values of m, a large sample approximation is generally used. Under the null hypothesis of no difference between the two sets of data, the expected value of W is

$$E(W) = n(m+n+1)/2, \text{ and the variance is}$$

$$Var(W) = mn(m+n+1)/12.$$

Then the distribution of $Z = \frac{W - E(W)}{Var(W)}$ tends toward a standard normal distribution and Table 1 page 19 can be used for testing for significance as described in Section III.

Let us examine the situation of ties more closely. First, we use average ranks to compute W. It is clear that if a tie occurs among two or more observations in the same group (e.g., "before" or "after" concentrations), it does not affect the value of W if average ranks are used for ties, or just assigned arbitrarily. For example, if the two concentrations of 42 in the "after" group had arbitrarily been ranked as 2 and 3, W would be the same. But, if ties occur among observations in different groups, W would change depending on how the ties were "broken." Averaging, as we did in the examples, is the usual tie-breaking procedure.

The variance of W is affected by ties, regardless of whether they are within or between groups. The variance, corrected for ties, is computed as follows:

$$Var_c(W) = \frac{mn}{12}\left[(m+n+1) - \frac{\sum_{j=1}^{g} t_j(t_j^2-1)}{(m+n)(m+n-1)}\right]$$

where g is the number of sets of ties and t_j is the size of tied set j. Then compute Z as above with $Var_c(W)$ replacing Var(W).

Although the large sample approximation is not applicable in our example, because m is only 10, we use it to demonstrate the computation of the correction for ties. We observe:

2 values of 42	(ties within "after" group)
2 values of 59	(ties between groups)
2 values of 99	(ties between groups)

Thus g = 3 and $t_1 = 2$, $t_2 = 2$ and $t_3 = 2$. Compute:

$$\sum_{j=1}^{g} t_j(t_j^2-1) = 2(2^2-1) + 2(2^2-1) + 2(2^2-1) = 18$$

and

$$Var_c(W) = \frac{(8)(10)}{12}\left[(8+10+1) - \frac{18}{(8+10)(8+10+1)}\right] = 126.27 .$$

The Wilcoxon rank sum test is sometimes referred to as the Wilcoxon two-sample test. An analogous test is the Mann-Whitney test which is based on an equivalent test statistic, U. In the case of no ties between concentration values, U = W - n (n + 1)/2, and therefore tests based on W and U are equivalent. One-tailed probabilities for U can be found in Siegel (1956), pp, 271-277. If there are ties, a correction is necessary in the computation of the Mann-Whitney U-statistic. For details see Hollander and Wolfe (1973) and Siegel (1956).

The Wilcoxon rank sum test can be performed using SAS. The procedure PROC NPAR1WAY with the Wilcoxon option may be used (see SAS User's Guide: Statistics, p. 205); an example output with the appropriate SAS statements is presented in the appendix.

The above procedure demonstrated how to use the Wilcoxon rank sum test to test for the presence of a step trend. It is also of interest to estimate the magnitude of such a change. A method of estimation based on the Wilcoxon test is presented next.

The first step in the estimation procedure is to calculate the differences formed by subtracting each observation in the after group from each observation in the before group. Denote these differences by D_{ij}, where

$$D_{ij} = X_i - Y_j, \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m.$$

In the example, m = 10 and n = 8, so that a total of 80 differences must be calculated. Next, order the nm differences from least to greatest and take the median (the middle value of the ordered differences) as the point estimate of the difference in the two groups. This is the estimate of the change.

Since the "after" observations were subtracted from the "before" observations, if the point estimate is positive, the interpretation would be that the introduction of the pollution control measure resulted in a decrease in the concentration. Applying the calculations to the example data results in a point estimate of 29 µg/ℓ as the decrease in concentration that occurred when the new pollution control measure was introduced. This is significantly different from zero, as concluded by the test.

D. Seasonal Kendall's Test for Trend

This is a test procedure proposed by Hirsch et al. (1982) which uses a modified form of Kendall's tau. In brief, if there are several years of monthly data, Kendall's K (number of concordant-disconcordant pairs), presented earlier, is computed for each of the 12 months, and the 12 statistics are then combined to provide a single overall test for trend. This method is discussed by van Belle and Hughes (1982) where it is also compared to another distribution-free test for trend proposed by Farrel (1980) using an aligned rank order test of Sen (1968).

1. Rationale

Figure 7 below depicts a series of 8 years of monthly data recorded for a given water quality parameter.

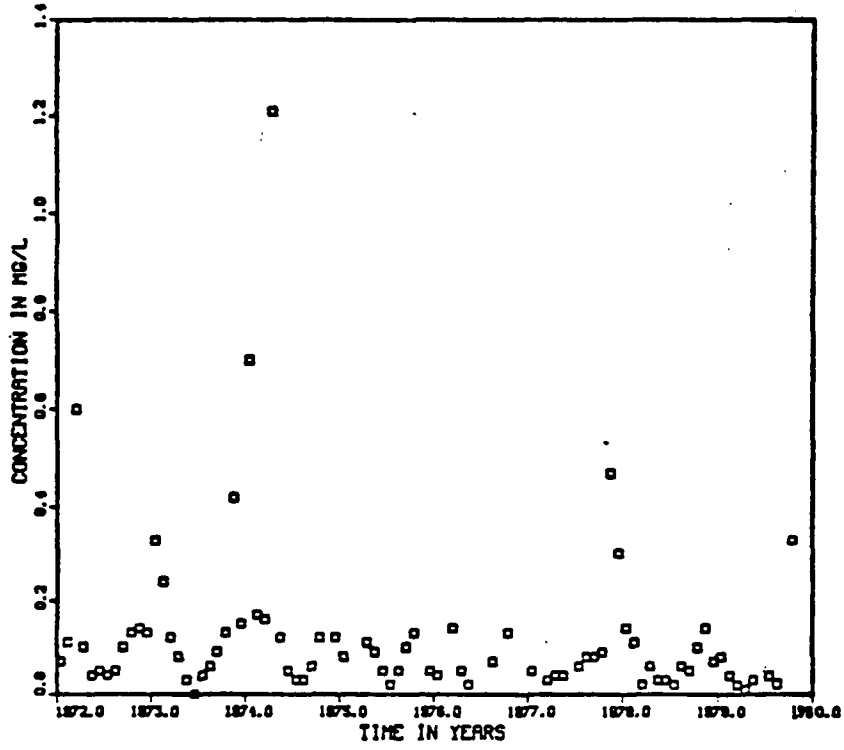


Figure 7. Monthly Concentrations of Total Phosphorus (ref. Hirsch et al., 1982)

The plot of the data clearly exhibits a seasonal movement--peaks and troughs recur at almost yearly intervals. This feature of having a period of a year (other periodicities may exist) is a common pattern for water quality parameters in general as discussed earlier. Comparing values between months within a year will thus not help in detecting a possible long-term trend over the time period considered. It would be more appropriate to make comparisons between data from the same month for different years, and avoid the problem of seasonality, and then combine the individual results into an overall test statistic from which we can draw conclusions about a trend.

2. Procedure

The following demonstration is based on monthly measurements; the same procedure can be applied to any sampling frequency (e.g., spring, summer, fall, winter measurements or average measurements), provided that the sampling scheme is identical from year to year. A general case with 12 months and n years of data is presented first; a simplified numerical example will then follow.

Arrange the monthly water quality measurements as follows:

		Month(j)						
		1	2	3	.	.	.	12
Year(i)	1	X_{11}	X_{12}	X_{13}	.	.	.	$X_{1,12}$
	2	X_{21}	X_{22}	X_{23}	.	.	.	$X_{2,12}$
	.	.						.
	.	.						.
	.	.						.
	n	X_{n1}	X_{n2}	.	.	.	.	$X_{n,12}$
Number of Observations		n_1	n_2	.	.	.	.	n_{12}

where X_{11} is the observation for the first month of the first year,
X_{21} is the observation for the first month of the second year,
$X_{2,12}$ is the observation for the 12th month of the 2nd year,
generally, X_{ij} is the observation for the jth month of the ith year.

Note that the number of observations need not be the same from month to month, i.e., there may be 5 January measurements ($n_1 = 5$), 6 February measurements ($n_2 = 6$), etc.

Next, make a second table of numbers of concordant (K+) and discordant (K-) pairs treating each month separately, using the procedure described with Kendall's tau. When finished, we have the following array:

	Month												
	1	2	3	.	.	.	.	.	12	number of:
K+	K_1^+	K_2^+	K_3^+	.	.	.	.	.	K_{12}^+	concordant pairs
K-	K_1^-	K_2^-	K_3^-	.	.	.	.	.	K_{12}^-	disconcordant pairs
K	K_1	K_2	K_3	.	.	.	.	.	K_{12}	(K+) - (K-)

Then sum the 12 monthly statistics to obtain $K = K_1 + K_2 + \ldots + K_{12}$. If the sample measurements are truly random (no trend), this statistic has a mean of 0 and a variance $Var(K) = Var(K_1) + \ldots + Var(K_{12})$. The variance of each monthly statistic K_j is computed as:

$$Var(K_j) = [n_j(n_j-1)(2n_j+5) - \sum_{t_i} t_i(t_i-1)(2t_i+5)]/18 \quad ,$$

where t_i is the size of the ith set of ties in the jth month.

Then compute the standard normal deviate Z with a continuity correction of one unit as:

$$Z = \begin{cases} \dfrac{K-1}{\sqrt{Var\ K}} & \text{if } K > 0 \\ 0 & \text{if } K = 0 \\ \dfrac{K+1}{\sqrt{Var\ K}} & \text{if } K < 0 \end{cases}$$

and use Table 1, page 19 of the standard normal deviates to determine the significance of Z.

Hirsch et al. (1982) have shown that the normal approximation works quite well with as few as 3 years of complete data. For fewer years of records, the exact distribution of $K_1 \ldots, K_{12}$, and therefore of K, has been derived by Kendall (1975).

3. Numerical Example

The following is an oversimplified example used for demonstration purposes only. For simplicity, we will assume that there were only four observations per year, taken on a quarterly basis. Note that no data were available in the first quarter of the fourth year. The example shows how such missing data may be handled. Consider the array of percent violations of a water quality standard for a given parameter:

		Quarter(j)			
		1	2	3	4
Year(i)	1	12.3	11.5	11.6	15.3
	2	10.5	10.9	9.5	14.8
	3	12.6	10.9	9.5	13.9
	4	-	9.8	9.5	14.0
Number of Observations		3	4	4	4

These data are plotted in Figure 8 below.

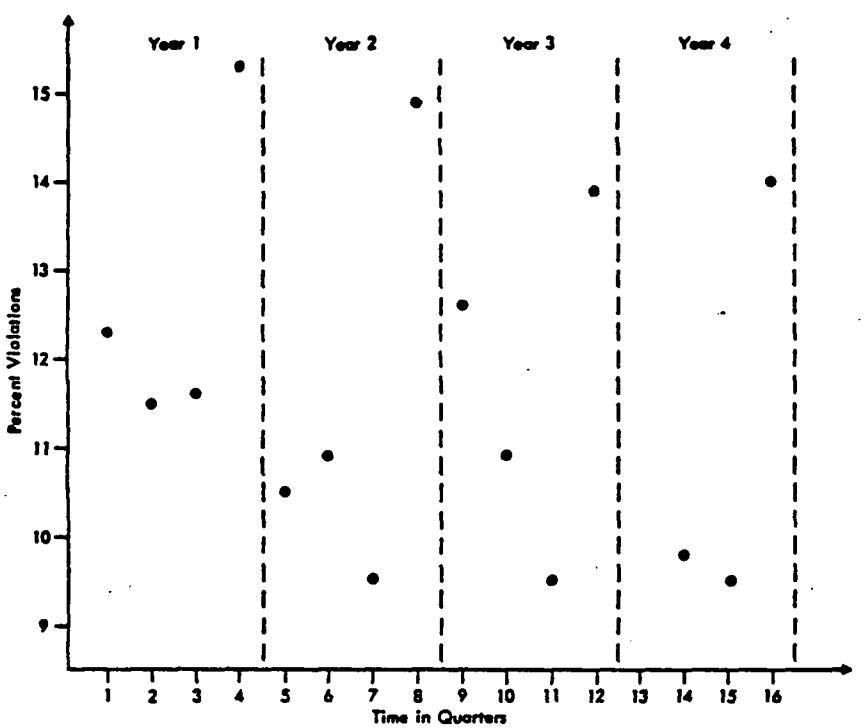


Figure 8. Plot of Percent Violations Versus Time

From the data we calculate the number of concordant and disconcordant pairs within each quarter, again ignoring ties in the computation of each K+ and K-. We obtain

	Quarter(j)			
	1	2	3	4
K+	2	0	0	1
K-	1	5	3	5
K=(K+)-(K-)	1	-5	-3	-4

Thus, $K_1 = 1$ and $n_1 = 3$; no ties
$K_2 = -5$ and $n_2 = 4$; 1 set of 2 ties (10.9, 10.9)
$K_2 = -3$ and $n_3 = 4$; 1 set of 3 ties (9.5, 9.5, 9.5)
$K_4 = -4$ and $n_4 = 4$; no ties,

and $K = K_1 + K_2 + K_3 + K_4 = -11$. The four variances, corrected for ties, are:

$\text{Var}(K_1) = [3(2)(11)-0]/18 = 66/18;$
$\text{Var}(K_2) = [4(3)(13)-2(1)(9)]/18 = 138/18;$
$\text{Var}(K_3) = [4(3)(13)-3(2)(11)]/18 = 90/18;$ and
$\text{Var}(K_4) = [4(3)(13)-0]/18 = 156/18$.

From here, $\text{Var}(K) = (66+138+90+156)/18 = 450/18 = 25$. Since K is negative, we compute Z as

$$Z = \frac{K+1}{\sqrt{\text{Var}(K)}} = \frac{-11+1}{\sqrt{25}} = -2 \ .$$

Since Z is less than -1.96, the lower critical Z corresponding to a two-sided 5% confidence level, (Table 1), we reject the null hypothesis of no trend in favor of the alternative that a trend is present.

Once we have identified a significant trend in a series of water quality measurements, we might be interested in determining the magnitude of the trend. For a set of stations at which trends have been detected, one could then compare the different trend slopes for a given water quality indicator and identify those stations where the trend slope is larger than average.

One way of computing the magnitude of a trend would be to compute the slope b, of the regression line of the water quality measurement versus time as we did earlier in Section IV. This technique, however, is recommended only with caution since the underlying assumptions for regression analysis are often violated when dealing with water quality measurements. A distribution-free method for computing the magnitude of a trend has been suggested by Hirsch et al. (1982). This method estimates the magnitude of trend by means of the seasonal Kendall slope estimator, B, computed as follows.

Considering again the more general data arrangement above, compute d_{ijk} quantities for each month (season) as follows:

$$d_{ijk} = (X_{jk} - X_{ik})/(j - i) \qquad \text{for all } (X_{jk}, X_{ik}) \text{ pairs}$$

where $k = 1,2. ., 12$ and $1 \leq i < j \leq n$. For monthly data, there will be a total of $12(n)(n-1)/2$ such differences. In general, with n years and m measurements per year, the number of differences will be $mn(n-1)/2$. The slope estimator, B, is the median of these d_{ijk} values (i.e., half the d_{ijk}'s exceed B and half fall short of it; if the number of differences is even, then take the average of the two middle ones). The estimator, B, is related to the seasonal Kendall test statistic S such that if S is positive, then B is positive or zero; if S is negative, then B is negative or zero. The S statistic is simply the number of positive d_{ijk} values minus the number of negative d_{ijk} values and B is the median of these d_{ijk}'s.

As a computation example for the second quarter ($j = 2$) and 4 years of data ($n = 4$), from the 4 measurements X_{12}, X_{22}, X_{32}, X_{42} compute the differences:

$$d_{122} = (X_{22} - X_{12})/1 = (10.9 - 11.5)/1 = -0.6$$

$$d_{132} = (X_{32} - X_{12})/2 = (10.9 - 11.5)/2 = -0.3$$

$$d_{142} - (X_{42} - X_{12})/3 = (9.8 - 11.5)/3 = -0.57$$

$$d_{232} = (X_{32} - X_{22})/1 = (10.9 - 10.9)/1 = 0$$

$$d_{242} = (X_{42} - X_{22})/2 = (9.8 - 10.9)/2 = -0.55$$

$$d_{342} = (X_{42} - X_{32})/1 = (9.8 - 10.9)/1 = -1.1$$

Note that for the first quarter, there will be only 3 (=(3)(2)/2) differences since one year had missing data, while for each of the remaining three quarters, there will be (4)(3)/2 = 6 differences.

Continuing the above calculations for all 21 pairs for the example data, one finds that the median of the differences is -0.5. Thus, we would estimate the trend as a decrease of 0.5 percent violations per quarter. Recall that this was found to be significantly different from zero at the two-sided 5% level using the seasonal Kendall's test.

The value, B, as a measure of trend magnitude, is quite resistant to the effect of extreme values in the data, unlike the slope of the regression line as computed in Section IV. It is also unaffected by seasonality because the slope is always computed between values that are multiples of m months (e.g., 12 months) apart.

A discussion of the seasonal Kendall's test for trend and other statistical procedures applied to total phosphorus measurements at NASQAN stations has been published by Smith et al. (1982). This document also contains a FORTRAN subroutine to perform the seasonal Kendall procedures.

E. Aligned Rank Sum Test for Seasonal Data (Step Trend)

This test is a method for testing for a step trend when the data exibit seasonality. This seasonal effect is usually clearly visible after the data have been plotted. In some cases, the analyst knows or suspects that a specific water quality parameter may be affected by seasonality.

When data are seasonal, the Wilcoxon rank sum test must be modified before testing for a step trend. Again, assume that a discrete event such as the opening of a new factory or the installation of a new pollution control system has been identified. The question is whether this event has produced a significant change in some measurement of a water quality parameter. The following is an outline of the computation of the appropriate distribution-free test statistic.

Assume measurements of a water quality indicator have been taken monthly at a fixed station over several years. More generally, the data can be collected for m seasons per year. The m times n measurements, where m is the number of months (seasons) and n is the number of years of collection, can be arranged as follows:

	1	2	3	.	.	.	m	mean
1	X_{11}	X_{12}	X_{13}	.	.	.	X_{1m}	$X_{1.}$
2	X_{21}	X_{22}	X_{23}*	.	.	.	X_{2m}	$X_{2.}$
.	.						.	.
Year .	.						.	.
.	.						.	.
mean	$X_{.1}$	$X_{.2}$	$X_{.3}$	.	.	.	$X_{.m}$	$X_{..}$

The * indicates the time at which a major event has happened.

For example, X_{21} is the measurement in the first month (season) of the second year. The symbol $X_{2.}$ denotes the average monthly measurement in the second year, while $X_{.1}$ would be the average of the January measurements over the n years.

Procedure

1. Within each month (column) subtract the monthly average from each measurement in the n years. This will result in an array of deseasonalized data. For example, in January, calculate the n differences $(X_{11} - X_{.1})$, $(X_{21} - X_{.1})$, ..., $(X_{n1} - X_{.1})$. These monthly differences will then have an average value of zero.

2. Rank all the nm differences from 1 to nm, regardless of month and year; this will produce the matrix of aligned ranks:

		1	2	3	.	.	.	m	mean
Year	1	R_{11}	R_{12}	R_{13}	.	.	.	R_{1m}	$R_{1.}$
	2	R_{21}	R_{22}	R_{23*}	.	.	.	R_{2m}	$R_{2.}$
	.	.						.	.
	.	.						.	.
	n	R_{n1}	.	.	.	.	.	R_{nm}	R_{n}
	mean	$R_{.1}$	.	.	.	.	.	$R_{.m}$	R..

3. Now sum the ranks of all the observations taken before the event in question. Let this sum of ranks be W. To construct the test we will use the fact that for large sample sizes the distribution of W will be approximately normal. Let b_i be the number of observations from month i that occurred before the change event, and let a_i be the number after the change event. In the example, $b_1=2$, $b_2=2$, $b_3=1$, ..., $b_m=1$, considering that the event occurred in the third month of the second year.

4. Next calculate

$$E = \sum_{i=1}^{m} b_i R_{.i} .$$

E is the sum of the products of the number of pre-event observations in each month and the mean rank of that month. E is the expected value of W.

5. Now calculate

$$V = \sum_{i=1}^{m} \frac{a_i b_i}{n_i(n_i-1)} \sum_{j=1}^{n_i} (R_{ij}-R_{.j})^2 ,$$

where $n_i = a_i+b_i$ is the number of observations for month i. If all months have the same number of observations this is n. V is the variance of W.

6. Then the test is based on:

$$Z = \frac{W-E}{\sqrt{V}} ,$$

which is tested using Table 1, (page 19).

Consider again the example used in Kendall's seasonal test. Assume that the measurements are taken quarterly rather than monthly (for simplicity of illustration) and that the event in question occurred with the beginning of the third quarter of the second year (denoted by *). Note that no observation was available for the first quarter of the fourth year. This example illustrates that the procedure can be used when there are missing data and shows how to apply the procedure in this case.

		Quarter			
		1	2	3	4
Year	1	12.3	11.5	11.6	15.3
	2	10.5	10.9	9.5*	14.8
	3	12.6	10.9	9.5	13.9
	4	-	9.8	9.5	14.0
Mean		11.8	10.8	10.0	14.5

Next, "deseasonalize" the data by subtracting from each observation within a quarter the mean value for this quarter. We obtain:

		Quarter			
		1	2	3	4
Year	1	0.50	0.72	1.57	0.80
	2	-1.30	0.12	-0.53	0.30
	3	0.80	0.12	-0.53	-0.60
	4	-	-0.98	-0.53	-0.50
Mean		0	-0.02	-0.02	0

Note: The means are not all exactly zero due to rounding errors.

The 15 new observations are then ranked and quarterly mean ranks are computed. Again, average ranks are used to break ties. The table of aligned ranks is as follows:

		Quarter			
		1	2	3	4
Year	1	11	12	15	13.5
	2	1	8.5	5	10
	3	13.5	8.5	5	3
	4	-	2	5	7
Mean		8.5	7.75	7.5	8.38

W is the sum of ranks over all four quarters of year one and the first two quarters of year two. (W = 11 + 12 + 15 + 13.5 + 1 + 8.5 = 61). The expected value of W, assuming no change, is the sum of the average ranks over the same period (E = 8.5 + 7.75 + 7.5 + 8.38 + 8.5 + 7.75 = 48.38). The variance consists of four terms, one for each quarter. Each term is the variance of the ranks within that quarter. For example, for the first quarter $b_1=2$, $a_1=1$, $n_1=3$, and the variance of the ranks is $(11-8.5)^2 + (1-8.5)^2 + (13.5-8.5)^2 = 87.5$. So the first term in the variance, corresponding to i=1, is:

$$\frac{2(1)}{(3)(2)} 87.5 = 29.17 .$$

Proceeding in the same way for the next three quarters and summing gives V=80.26. Then

$$Z = \frac{61 - 48.38}{\sqrt{80.26}} = 1.41 .$$

Comparing this value of Z to the critical value of 1.96 at the two-sided 5% level (Table 1), shows that the change is not significantly different from zero because 1.41 falls between -1.96 and +1.96.

The positive sign for Z indicates that the "before" period had higher ranks (after deseasonalizing) than the "after" period, thus, the direction of change was from high to low. However, in this case the change could be due to random fluctuations.

More details on the aligned rank sum test can be found in Lehmann (1975), p. 132-141. Unfortunately, this procedure is not available through SAS.

F. Trend and Change

It is very difficult in nonparametric statistics to deal with a data set containing both a step trend and a long-term trend, or to distinguish between the two trends in a series of data values. In general one should only be interested in a step trend when there is a definite external event that would be likely to result in a change. That is, a step trend is only present when an event occurred at a known point in time and influenced the data. The testing procedures for change could be misled by the presence of a long-term trend. Likewise the tests for trend could indicate an apparent trend in the presence of (only) a step trend. If there is a change event, determining whether it is significant and whether there is a trend as well, or only one, is a difficult task. The importance of plotting the data must be re-emphasized.

In the parametric case, as discussed in Section IV, one can use multiple regression analysis to test for the presence of both a step trend and a long-term trend in the same data series. Unfortunately, there is no distribution-free procedure that is as well developed and as easy to apply. One could try both types of tests (for change and trend). If one is significant and the other not, then the answer is reasonably clear. If neither is significant, then the data appear to be random. However, if both are significant, then both types of nonrandomness may be present, or only one. To determine whether both a change and a trend are present or only one, and if only one, which, will require a series of analyses, and advice from a statistician should be sought.

To test for a long-term trend in the presence of a step trend, one could use the rank procedure as in subsection D. The "before" and "after" data would be considered as two groups and separate means calculated. Differences

between each observation and its group mean would be calculated and the Kendall's test applied. In effect this would be the seasonal Kendall's test with only two "seasons"--before and after. To test for the presence of a step trend when a long-term trend is known to exist would require estimating the slope and calculating the difference between each observation and the trend line, then applying the Wilcoxon rank-sum test to the differences in a manner similar to the procedure of Subsection E. A detailed presentation of these procedures is beyond the scope of this document.

VI. SPECIAL PROBLEMS

As mentioned throughout the preceding sections, water quality data do not, in general, exhibit all the desired properties necessary for the use of parametric statistical procedures. We have already mentioned the fact that most water quality measurements show seasonal (or cyclical) effects. We have then suggested methods to deseasonalize the data when using parametric procedures (i.e., multiple regression) or distribution-free methods (i.e., seasonal Kendall's test, aligned rank sum test). (A seasonal adjustment method has also been proposed in Schlicht (1981), although within the more general setting of time series analysis.)

Other problems inherent to "real life" data bases are those of missing data (incomplete records) and extreme or outlying observations. Another fact mentioned throughout Section IV is the assumption of normality of the deseasonalized data or residuals obtained from regression analysis. In addition, a problem specific to water quality measurements, especially when concentrations of pesticides, trace metals, etc. are estimated, are measurements below detection limit. One final and important point is the problem of flow changes in rivers and streams which will affect concentrations of most constituents considered as potential water quality indicators.

A. Missing Data

The basic assumption we make in treating missing data, from a statistical point of view, is that data are missing because of mistakes, such as lost records, and not because the analyst simply wants to ignore conflicting or compromising data. In other words, any missing datum is assumed to follow the same pattern as the recorded observations. A simple method to fill in a few missing data values is to replace them by the sample mean. More involved methods deal with least squares estimates; these methods are available through standard statistical program packages such as SAS, BMDP and SPSS, and are explained, for example, in Johnson and Leone (1977).

The application of the seasonal Kendall's test for trend is not restricted to complete data sets, nor is it necessary to have full years of data. As we mentioned earlier, the seasonal Kendall's test statistic can in fact be computed with incomplete data. It is also suggested (van Belle and Hughes (1982)) that the season length be adjusted, if necessary, to obtain a reasonable record within each season. When using parametric procedures, a few missing data will reduce the sample size and slightly affect the means and standard deviations. A large number of missing data, however, might affect these statistics considerably.

B. Outlying Observations

Most often, outlying or extreme observations, also called outliers, can be easily detected either when looking at a plot of the data or even earlier, when closely examining the raw data sheets. Several logical steps can be taken when outliers are found. The most basic first step is to double check suspicious observations for transcription errors. If the entry was an error, correct it if possible. If the proper entry cannot be retrieved but an error is certain, then delete the datum. There are also statistical procedures to test whether a suspected outlier(s) is in fact an extreme observation. Such methods are presented in detail in ASTM Standard E178-75 entitled "Standard Recommended Practice for Dealing With Outlying Observations." Also, tests for outliers based on ranks are presented in the "Handbook of Tables for Probability and Statistics" (1966).

Extreme observations can be actual and caused by flow changes, temperature changes, etc. It is therefore recommended that ancillary data such as time of day, water temperature and rate of discharge at the time of sample collection be collected at the same time as the water quality data. Many outliers can then be explained and/or corrected. Most rank tests, as described earlier, are little affected by the magnitude of the observations and therefore by outliers. In parametric tests, however, outliers affect means and variances, and may therefore invalidate the resulting tests and conclusions.

C. Test for Normality

Given a set of measurements of a variable X, one wishes to know whether the variable comes from a normal distribution. The following simple plotting procedure, if the data base is not too extensive, can be used. Consider the following example of 12 data points. These n=12 data points can be rearranged in ascending order:

i	X	(i/n+1)x100%
1	-1.45	7.7
2	-1.35	15.4
3	-0.78	23.1
4	-0.62	30.8
5	-0.01	38.5
6	0.04	46.2
7	0.22	53.8
8	0.49	61.5
9	0.72	69.2
10	1.45	76.9
11	1.79	84.6
12	2.50	92.3

Should a value of X occur more than once, then the corresponding value of i (cumulative frequency) increases appropriately. The maximum value of i is always n, the total number of data points. The pairs (X, (i/n+1)100)-values are then plotted on probability paper using an appropriate scale for X on the horizontal axis. Figure 9 shows the results. The vertical axis for the values of (i/n+1)x100% is already scaled from 0.01 to 99.99. If the data came from a normal distribution, then the plotted points would fall on a straight line. In practice, a straight line can be drawn by hand through the points and a judgment can be made as to the normality of the data. Also, rough estimates of the mean, $\bar{X}$, and standard deviation, S, can be made from this plot. The horizontal line drawn through 50 cuts the plotted line at the value of $\bar{X}$, and the horizontal line through 84 cuts it at the value of $\bar{X}$ + S; these two numbers then yield the value of S by subtraction.

More rigorous statistical tests are available for testing for normality, such as the Kolmogorov-Smirnov test (see Hollander and Wolfe, 1973) or the χ^2 goodness of fit test (see Snedecor and Cochran, 1980). The drawback of using these tests is that they tend to easily reject the hypothesis

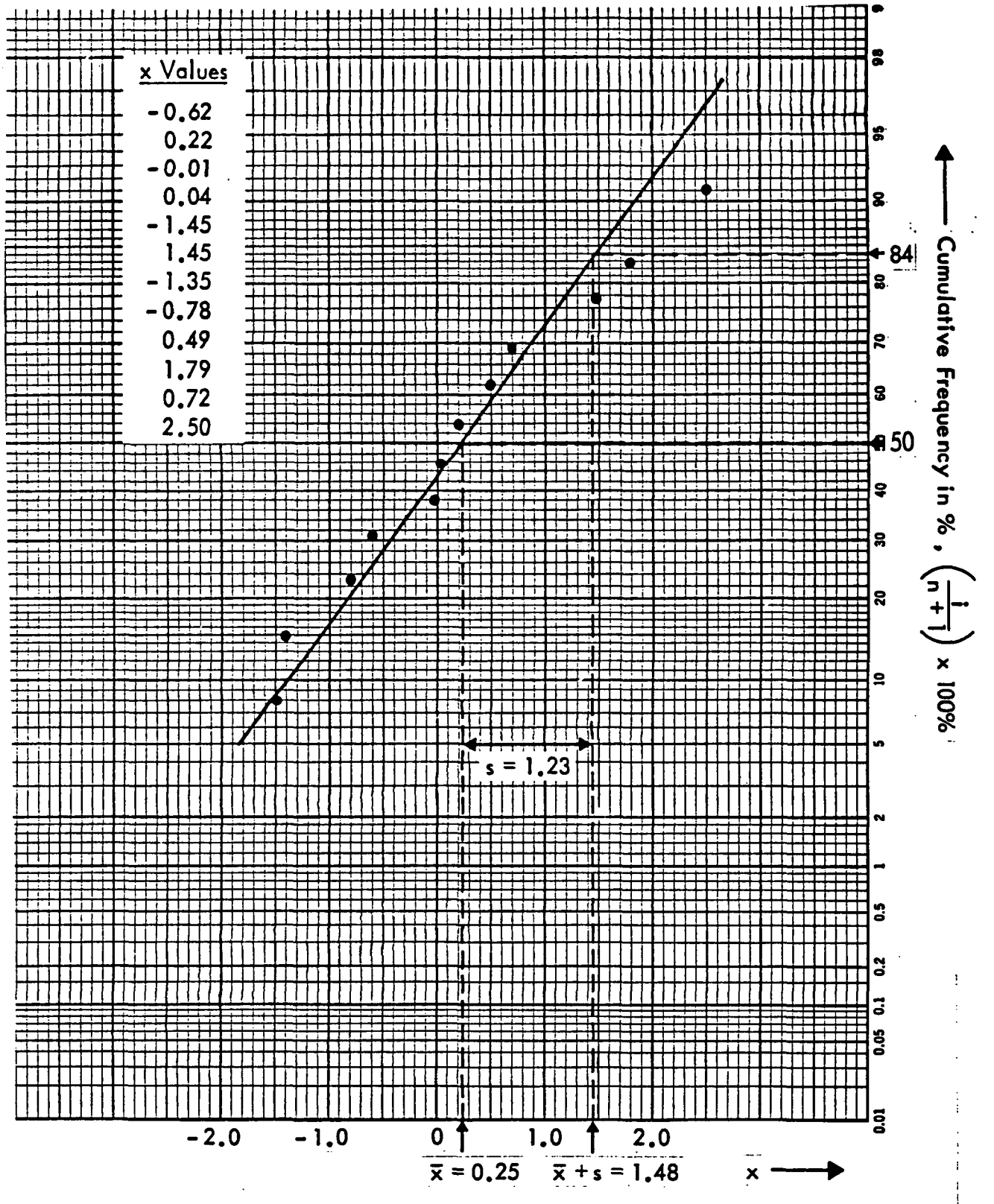


Figure 9. Example of plot on probability paper.

of normality when the sample sizes are fairly large, although a visual examination of a probability plot tends to support the hypothesis. On the other hand, when sample sizes are relatively small, as might be the case when only two-years worth of data are available, then the more rigorous tests tend not to detect deviations from normality. Thus, a probability plot might be more helpful. It should be noted here that SAS provides a probability plot, as well as a test for normality, through the UNIVARIATE procedure (see example output in the appendix).

D. Detection Limits

Sometimes in testing for metals, organic compounds, etc., the analytical test will not be sensitive enough to quantify the amount of the particular substance. The amount is below the detection limit of the test, so is reported simply as "less than" that limit. These data, also referred to as "censored" data, are not to be confused with missing data, because they are not "missing"--they are known to be in a certain range, but the precise values are not known.

There are, however, some ways to compute values that can be substituted for "less than" values. Such options would include deleting the sample, filling in with zeros, substituting the actual detection limit for the datum, or filling in with a random number between zero and the detection limit based on the underlying distribution of the data. The first three methods are all biased in some way. The last one, on the other hand, requires information about the distribution. A simple random number between zero and the detection limit is sometimes used for substitution, implying a uniform distribution.

Rank tests can handle "less than" values in some cases with less difficulty than parametric tests. If the limit of detection is constant, all "less than" values for a particular constituent are considered as "ties." Alternatively, rank tests that treat censored data explicitly have been developed (e.g., Gehan 1965). Note that detection limit values for different constituents are handled individually.

E. Flow Adjustments

Caution must be exercised when interpreting trends found to be significant by any of the previously described statistical procedures, especially when the measurements used are specific constituent concentrations. It is common knowledge that for most constituents, concentrations change as the flow changes, which in turn introduces considerable variability into the measurements. Flow conditions can vary naturally due to climatic factors, and artificially due to stream regulation and manipulation by man.

One way to correct for changing flow is to determine the relationship between flow and concentration of the considered constituent. However, no uniform equation exists since this relationship may vary from site to site and from constituent to constituent. Hirsch et al. (1980) suggested some nonlinear equations characterizing relationships between concentrations and flow in cases where the increased discharge of a constituent is due to precipitation, snowmelt, or reservoir release. In another case, quadratic equations are proposed to relate concentrations and flow when the constituent load may increase dramatically with an increase in discharge because of runoff during a storm event.

When the effect of increased discharge is a simple dilution effect, the relationship between concentration and discharge can be characterized by

$$X = \lambda_1 + \lambda_2/Q \text{ , or}$$

$$X = \lambda_1 + [\lambda_2/(1 + \lambda_3 Q)] \text{ , for example,}$$

where X is the concentration, Q is the discharge flow, and the coefficients λ_1 and λ_2 are equal to or greater than zero, and λ_3 is greater than zero. Generally the coefficients in these equations can be estimated via least squares methods (e.g., regression analysis).

The sequence of procedures suggested by Hirsch et al. (1980) can be summarized as follows:

1. First find the best fitting relationship between flow and concentration using regression methods.

2. Compute the series of flow-adjusted concentrations whenever the relationship determined in Step 1 is significant.

3. Apply the seasonal Kendall test for trend.

4. Compute the magnitude of the trend, if significant, using the seasonal Kendall slope estimator.

For an example of these procedures, the interested reader is referred to Smith et al. (1982), who applied all three methods--seasonal Kendall test for trend, flow adjustment, and seasonal Kendall slope estimator--to measurements of total phosphorus concentrations.

BIBLIOGRAPHY

ASTM Designation: E178-75. 1975 "Standard Recommended Practice For Dealing With Outlying Observations."

Bell, Charles B., and E. P. Smith. 1981. Water Quality Trends: Inference for First-Order Autoregressive Schemes, Tech. Rep. 6, SIAM Instit. for Math. in Soc., Biomath. Group, Univ. of Wash., Seattle.

Box, George E. P., and J. M. Jenkins. 1970. Time Series Analysis. Holden-Day, San Francisco, Ca.

Box, George E. P., and G. C. Tiao. 1975. "Intervention Analysis with Application to Economic and Environmental Problems." J. American Statistical Assoc., Vol. 70, pp. 70-79.

Chatterjee, Samprit, and B. Price. 1977. Regression Analysis by Example. John Wiley and Sons, New York.

Draper, Norman R., and H. Smith. 1981. Applied Regression Analysis. Second Edition, John Wiley and Sons, Inc., New York.

Dykstra, Richard L. and T. Robertson. 1983. "On Testing Monotone Tendencies." J. American Statistical Assoc., Vol. 78, pp. 342-350.

Farrell, Robert L. 1980. Methods for Classifying Changes in Environmental Conditions. Tech. Rep. VRI-EPA7.4-FR80-1, Vector Research, Inc., Ann Arbor, Mich.

Gehan, Edmund A. 1965. "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly Censored Samples." Biometrika 521, 203-223.

General Accounting Office. 1981. Better Monitoring Techniques are Needed to Assess the Quality of Rivers and Streams. Report CED-81-30, U.S. General Accounting Office, Washington, D.C.

Colorado.

Handbook of Tables for Probability and Statistics. 1966. Edited by Beyer, William H. The Chemical Rubber Co.

Hirsch, Robert M., J. R. Slack, and R. A. Smith. 1982. "Techniques of Trend Analysis for Monthly Water Quality Data." Water Resources Research, Vol. 18(1), pp. 107-121.

Hollander, Myles, and D. A. Wolfe. 1973. Nonparametric Statistical Methods. John Wiley and Sons, New York.

Jernigan, Robert W., and J. C. Turner. Seasonal Trends in Unequally Spaced Data: Confidence Intervals for Spectral Estimates. Submitted for publication.

Johnson, Norman L, and F. C. Leone. 1977. 2 Vol. Statistics and Experimental Design in Engineering and the Physical Sciences. Second Edition, John Wiley and Sons, Inc., New York.

Kendall, Maurice G., and W. R. Buckland. 1971. A Dictionary of Statistical Terms. Third Edition. Hafner Publishing Company, Inc., New York.

Kendall, Maurice G., and A. Stuart. 1966. The Advanced Theory of Statistics, Volume 3. Hafner Publ. Co., New York, pp. 342.

Kendall, Maurice G. 1975. Rank Correlation Methods. Charles Griffin, London.

Langley, Russell A. 1971. Practical Statistics Simply Explained. Second Edition, Dover Publications, Inc., New York.

Lehmann, Erich L. 1975. Nonparametric Statistical Methods Based on Ranks. Holsten Day, San Francisco.

Lettenmaier, Dennis P. 1976. "Detection of Trends in Stream Quality: Monitoring Network Design and Data Analysis." Tech. Rep. 51, Harris Hydraul. Lab., Dept. of Civil. Eng., Univ. of Wash., Seattle.

Lettenmaier, Dennis P. 1976. "Detection of Trends in Water Quality Data from Records With Dependent Observations." Water Resources Research, Vol. 12(5), pp. 1037-1046.

Mann, Henry B. 1945. "Nonparametric Tests Against Trend." Econometrica, Vol. 13, pp. 245-259.

Sen, Pranab K. 1968. "On A Class of Aligned Rank Order Tests in Two-Way Layouts." Annals of Mathematical Statistics, Vol. 39, pp. 1115-1124.

Schlicht, Ekkehart. 1981. "A Seasonal Adjustment Principle and a Seasonal Adjustment Method Derived from this Principle." Journal of the American Statistical Association, Vol. 76, pp. 374-378.

Siegel, Sidney. 1956. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York.

Smith, Richard A., R. M. Hirsch, and J. R. Slack. 1982. A Study of Trends in Total Phosphorus Measurements at NASQAN Stations. U.S. Geological Survey Water-Supply Paper 2190.

Snedecor, George W., and W. G. Cochran. 1980. Statistical Methods, Seventh Edition, the Iowa State University Press, Ames, Iowa.

STORET User Handbook. 1980. U.S. Environmental Protection Agency, Office of Water and Hazardous Materials.

van Belle, Gerald, and J. P. Hughes. 1982. "Nonparametric Tests for Trend in Water Quality." SIMS Technical Report No. 11, University of Washington, Seattle, Washington (to appear in Water Resources Research).

van Belle, Gerald, and J. P. Hughes. 1983. "Monitoring for Water Quality: Fixed Station versus Intensive Surveys." Journal Water Pollution Control Federation. Vol. 55, pp. 400-404.

Wallis, W. Allen, and H. V. Roberts. 1956. Statistics: A New Approach. The Free Press, New York.

Statistical Program Packages:

BMDP Statistical Software, 1980, University of California Press

SAS: Statistical Analysis System, SAS Institute, Inc.,
SAS User's Guide: Basics. 1982 Edition
SAS User's Guide: Statistics. 1982 Edition
Box 8000, Cary, North Carolina

SPSS: Statistical Package for the Social Sciences, 1982, McGraw-Hill.

APPENDIX

This appendix is a collection of output examples using SAS and the data used to demonstrate various procedures in the preceding sections. It is organized in the same fashion as the text, and the title in each output corresponds to the appropriate section and subsection. The following procedures are shown:

1. Regression analysis (IV-B)
2. Student's t-test (IV-C)
3. Multiple regression analysis (IV-D)
4. Kendall's tau test (V-B)
5. Wilcoxon rank sum test (V-C)

A FORTRAN subroutine for the seasonal Kendall's test and trend magnitude estimation is included in Smith et al. (1982).

1. REGRESSION ANALYSIS (IV-B)

```
① OPTIONS LINESIZE=100 NODATE ;
  DATA REGRESS;
  INPUT YEAR WQI @@ ;
② CARDS;
  1977 46 1978 52 1979 42 1980 44 1981 39 1982 45 1983 40
  TITLE REGRESSION ANALYSIS , WQI VERSUS TIME(YR);
  TITLE3 OUTPUT EXAMPLE FOR SECTION IV B;
③ PROC REG ;
  MODEL WQI=YEAR ;
  OUTPUT OUT=RES
  PREDICTED=PRED
  RESIDUAL=RESID ;
④ PROC PLOT DATA=RES;
  PLOT PRED*YEAR='P'
       WQI*YEAR='O'/OVERLAY ;
⑤ PROC PRINT DATA=RES ;
⑥ PROC UNIVARIATE DATA=RES PLOT NORMAL;
  VAR RESID;
  TITLE5 TEST OF NORMALITY FOR THE RESIDUALS;
  /*
```

① Data set up

② Data (if data are on file, use an Infile statement instead)

③ Regression procedure

④ Plotting procedure

⑤ Printing procedure

⑥ Univariate procedure with plot and normality test options.

Titles are optional

REGRESSION ANALYSIS , WQI VERSUS TIME(YR)

OUTPUT EXAMPLE FOR SECTION IV B

DEP VARIABLE: WQI

(sample size − 1)

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	1	43.750000	43.750000	3.114	0.1379
ERROR	5	70.250000	14.050000		
C TOTAL	6	114.000			

Significance of the model; if less than α (e.g., 0.05) then the model is significant; this model is Not significant at the 5% level

ROOT MSE	3.748333	R-SQUARE	0.3838
DEP MEAN	44.000000	ADJ R-SQ	0.2605
C.V.	8.518939		

$= 0.620^2 = r^2$

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	2519.000	1402.570	1.796	a) 0.1324
YEAR = slope	1	-1.250000	0.708368	-1.765	b) 0.1379

two-sided levels of significance for
a) intercept = 0
b) slope = 0

if these probabilities are higher than α (e.g. 0.05) then the parameters are not significantly different from 0.

the regression equation is: $Y = -1.25\ \text{Year} + 2519$ with year = 1977, 78 ect

Note: if Year No were used (1977 = 1, 1978 = 2, etc) then

$$Y = -1.25 \cdot \text{Year No} + (2519 - 1.25 \cdot 1976)$$

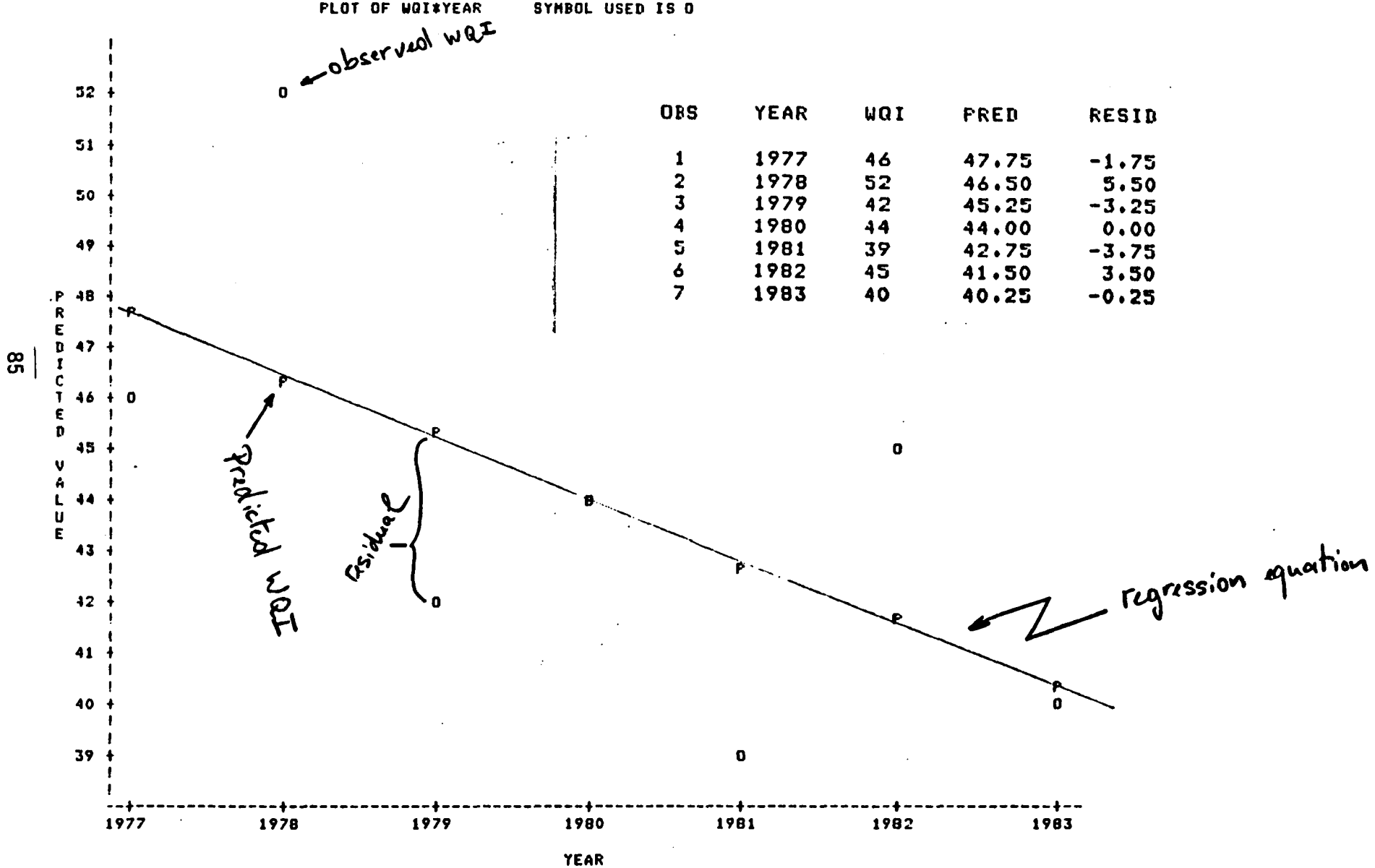
or $Y = -1.25\ \text{Year No} + 49$

Results: the regression model is not significant at the 5% level

OUTPUT EXAMPLE FOR SECTION IV B

PLOT OF PRED*YEAR SYMBOL USED IS P
PLOT OF WQI*YEAR SYMBOL USED IS O

OBS	YEAR	WQI	PRED	RESID
1	1977	46	47.75	-1.75
2	1978	52	46.50	5.50
3	1979	42	45.25	-3.25
4	1980	44	44.00	0.00
5	1981	39	42.75	-3.75
6	1982	45	41.50	3.50
7	1983	40	40.25	-0.25



Variable on which test is performed

```
                    REGRESSION ANALYSIS , WQI VERSUS TIME(YR)

                        OUTPUT EXAMPLE FOR SECTION IV B

                      TEST OF NORMALITY FOR THE RESIDUALS

                                  UNIVARIATE

VARIABLE=RESID          RESIDUALS

                MOMENTS                                QUANTILES(DEF=4)                      EXTREMES

N                  7  SUM WGTS             7  100% MAX        5.5      99%        5.5     LOWEST    HIGHEST
MEAN       1.787E-13  SUM          1.251E-12   75% Q3         3.5      95%        5.5      -3.75      -1.75
STD DEV      3.42174  VARIANCE       11.7083   50% MED      -0.25      90%        5.5      -3.25      -0.25
SKEWNESS    0.680336  KURTOSIS     -0.689156   25% Q1       -3.25      10%      -3.75      -1.75  1.705E-13
USS            70.25  CSS              70.25    0% MIN      -3.75       5%      -3.75      -0.25        3.5
CV         1.915E+15  STD MEAN        1.2933                            1%      -3.75  1.705E-13        5.5
T:MEAN=0   1.381E-13  PROB>|T|             1   RANGE         9.25
SGN RANK          -1  PROB>|S|      0.932647   Q3-Q1         6.75
NUM ^= 0           7                           MODE         -3.75
W:NORMAL    0.923974  PROB<W           0.482

STEM LEAF                    #  BOXPLOT                   NORMAL PROBABILITY PLOT
   4 5                       1     |             5+                                            *
   2 5                       1  +-----+           |                                       *
   0                            |  +  |          1+                           +   *
  -0 730                     3  *-----*           |                       *   *
  -2 82                      2  +-----+         -3+           *     *
     ----+----+----+----+                         +----+----+----+----+----+----+----+----+----+----+
                                                      -2        -1        +0        +1        +2
```

if greater than α (e.g. 0.05) then accept hypothesis of normality

test of normality of residuals

- if N is less than or equal to 50, then W is computed
- if N is greater than 50, then a D statistic is computed

2. STUDENT'S T-TEST (IV-C)

```
①  OPTIONS LINESIZE=100 NODATE ;
   DATA WATERQ;
   INPUT TIME $ CONC @@;
   LABEL CONC=CONCENTRATION;
②  CARDS ;
   B 99 B 111 B 74 B 123 B 71 B 75 B 59 B 85
   A 59 A 99 A 82 A 51 A 48 A 39 A 42 A 42 A 47 A 50
   TITLE T-TEST PROCEDURE , TOTAL CHROMIUM CONCENTRATIONS;
   TITLE3 OUTPUT EXAMPLE TO SECTION IV C ;
③  PROC TTEST ;
   CLASS TIME;
   VAR CONC ;
④  PROC UNIVARIATE PLOT NORMAL ;
   VAR CONC ;
   TITLE5 TEST OF NORMALITY AND NORMAL PROBALITY PLOT ;
   /*
```

① Data setup

② Data (in the program); if the data are on file then this part is ignored and an Infile statement is used

③ t-test procedure

④ Univariate procedure with Plot and Normality test options

Titles are optional

T-TEST PROCEDURE , TOTAL CHROMIUM CONCENTRATIONS

OUTPUT EXAMPLE TO SECTION IV C

TTEST PROCEDURE

VARIABLE: CONC CONCENTRATION

TIME	N	MEAN	STD DEV	STD ERROR	VARIANCES	T	DF	PROB > |T|
A	10	55.90000000	19.49615347	6.16522506	UNEQUAL	-3.1503	14.2	① 0.0070
B	8	87.12500000	21.95083793	7.76079318	EQUAL	-3.1946	16.0	② 0.0056

FOR HO: VARIANCES ARE EQUAL, F'= 1.27 WITH 7 AND 9 DF PROB > F'= 0.7234

mean value of "after" data

mean value of "before" data

Last line: F-test for equality of variances

- if Prob is greater than α (e.g., 0.05) then accept null hypothesis of equal variance and read result for t-test in ②
- if Prob is less than or equal to α, then reject null hypothesis and read results of adjusted t-test in ①

two-sided levels of significant for t-test. If greater than α, accept null hypothesis that the means are equal, otherwise reject.

Results: the means are significantly different (95% level) by an amount of $(87.125 - 55.9) = 31.225$

T-TEST PROCEDURE , TOTAL CHROMIUM CONCENTRATIONS

OUTPUT EXAMPLE TO SECTION IV C

TEST OF NORMALITY AND NORMAL PROBALITY PLOT

UNIVARIATE

VARIABLE=CONC CONCENTRATION

MOMENTS

N	18	SUM WGTS	18
MEAN	69.7778	SUM	1256
STD DEV	25.5839	VARIANCE	654.536
SKEWNESS	0.653072	KURTOSIS	-0.612432
USS	98768	CSS	11127.1
CV	36.6648	STD MEAN	6.03018
T:MEAN=0	11.5714	PROB>|T|	0.0001
SGN RANK	85.5	PROB>|S|	.000212983
NUM ^= 0	18		
W:NORMAL	0.920973	PROB<W	0.166

QUANTILES(DEF=4)

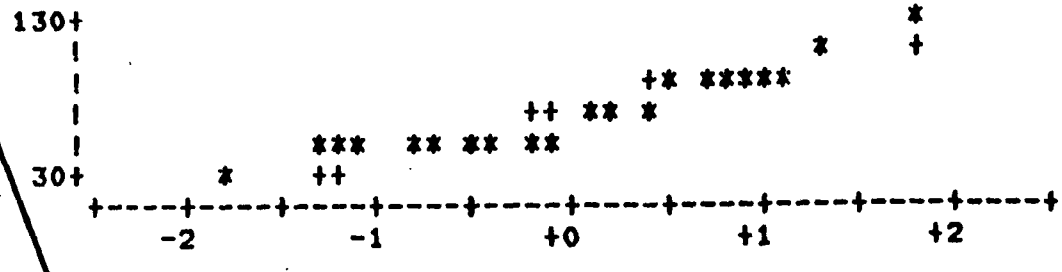
100% MAX	123	99%	123
75% Q3	88.5	95%	123
50% MED	65	90%	112.2
25% Q1	47.75	10%	41.7
0% MIN	39	5%	39
		1%	39
RANGE	84		
Q3-Q1	40.75		
MODE	42		

EXTREMES

LOWEST	HIGHEST
39	85
42	99
42	99
47	111
48	123

```
STEM LEAF                          #  BOXPLOT
  12 3                             1     |
  10 1                             1     |
   8 2599                          4 +-----+
   6 145                           3 *--+--*
   4 22780199                      8 +-----+
   2 9                             1     |
     ----+----+----+----+
     MULTIPLY STEM.LEAF BY 10**+01
```

NORMAL PROBABILITY PLOT



Test of normality. Since 0.166 is greater than α of 0.05, accept hypothesis that data come from a normal distribution

3. MULTIPLE REGRESSION ANALYSIS (IV-D)

```
OPTIONS LINESIZE=100 NODATE ;
DATA MULTIPLE;
INPUT TIME  CONC ;
IF MOD(TIME,12)=1 THEN X1=1 ;ELSE X1=0;
IF MOD(TIME,12)=2 THEN X2=1 ;ELSE X2=0;
IF MOD(TIME,12)=3 THEN X3=1 ;ELSE X3=0;
IF MOD(TIME,12)=4 THEN X4=1 ;ELSE X4=0;
IF MOD(TIME,12)=5 THEN X5=1 ;ELSE X5=0;
IF MOD(TIME,12)=6 THEN X6=1 ;ELSE X6=0;
IF MOD(TIME,12)=7 THEN X7=1 ;ELSE X7=0;
IF MOD(TIME,12)=8 THEN X8=1 ;ELSE X8=0;
IF MOD(TIME,12)=9 THEN X9=1 ;ELSE X9=0;
IF MOD(TIME,12)=10 THEN X10=1 ;ELSE X10=0;
IF MOD(TIME,12)=11 THEN X11=1 ;ELSE X11=0;
IF MOD(TIME,12)=0 THEN X12=1 ;ELSE X12=0;
IF TIME LE 18 THEN C=0 ; ELSE C=1 ;
CARDS;

TITLE MULTIPLE REGRESSION ANALYSIS;
TITLE3 OUTPUT EXAMPLE TO SECTION IV D ;
PROC PRINT ;
FORMAT CONC 5.3 ;
PROC REG ;
MODEL CONC=X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 C TIME/NOINT ;
PROC PLOT ; PLOT CONC*TIME=C/HAXIS=0 TO 72 BY 2 HREF=18;
/*
```

① Data setup, the IF statements define the dummy variables $X_1, \ldots, X_{12}$

② Print procedure (avoid if data set is too long!)

③ Regression procedure

MULTIPLE REGRESSION ANALYSIS 1

OUTPUT EXAMPLE TO SECTION IV D

OBS	TIME	CONC	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	C
1	1	0.704	1	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0.636	0	1	0	0	0	0	0	0	0	0	0	0	0
3	3	0.288	0	0	1	0	0	0	0	0	0	0	0	0	0
4	4	0.576	0	0	0	1	0	0	0	0	0	0	0	0	0
5	5	0.422	0	0	0	0	1	0	0	0	0	0	0	0	0
6	6	0.198	0	0	0	0	0	1	0	0	0	0	0	0	0
7	7	0.414	0	0	0	0	0	0	1	0	0	0	0	0	0
8	8	0.057	0	0	0	0	0	0	0	1	0	0	0	0	0
9	9	0.028	0	0	0	0	0	0	0	0	1	0	0	0	0
10	10	0.505	0	0	0	0	0	0	0	0	0	1	0	0	0
11	11	0.113	0	0	0	0	0	0	0	0	0	0	1	0	0
12	12	0.406	0	0	0	0	0	0	0	0	0	0	0	1	0
13	13	0.414	1	0	0	0	0	0	0	0	0	0	0	0	0
14	14	0.624	0	1	0	0	0	0	0	0	0	0	0	0	0
15	15	0.540	0	0	1	0	0	0	0	0	0	0	0	0	0
16	16	0.543	0	0	0	1	0	0	0	0	0	0	0	0	0
17	17	0.618	0	0	0	0	1	0	0	0	0	0	0	0	0
18	18	0.372	0	0	0	0	0	1	0	0	0	0	0	0	0
19	19	0.198	0	0	0	0	0	0	1	0	0	0	0	0	1
20	20	0.125	0	0	0	0	0	0	0	1	0	0	0	0	1
21	21	0.139	0	0	0	0	0	0	0	0	1	0	0	0	1
22	22	0.010	0	0	0	0	0	0	0	0	0	1	0	0	1
23	23	0.117	0	0	0	0	0	0	0	0	0	0	1	0	1
24	24	0.265	0	0	0	0	0	0	0	0	0	0	0	1	1
25	25	0.256	1	0	0	0	0	0	0	0	0	0	0	0	1
26	26	0.366	0	1	0	0	0	0	0	0	0	0	0	0	1
27	27	0.342	0	0	1	0	0	0	0	0	0	0	0	0	1
28	28	0.235	0	0	0	1	0	0	0	0	0	0	0	0	1
29	29	0.487	0	0	0	0	1	0	0	0	0	0	0	0	1
30	30	0.141	0	0	0	0	0	1	0	0	0	0	0	0	1
31	31	0.225	0	0	0	0	0	0	1	0	0	0	0	0	1
32	32	0.010	0	0	0	0	0	0	0	1	0	0	0	0	1
33	33	0.124	0	0	0	0	0	0	0	0	1	0	0	0	1
34	34	0.019	0	0	0	0	0	0	0	0	0	1	0	0	1
35	35	0.133	0	0	0	0	0	0	0	0	0	0	1	0	1
36	36	0.080	0	0	0	0	0	0	0	0	0	0	0	1	1
37	37	0.296	1	0	0	0	0	0	0	0	0	0	0	0	1
38	38	0.370	0	1	0	0	0	0	0	0	0	0	0	0	1
39	39	0.328	0	0	1	0	0	0	0	0	0	0	0	0	1
40	40	0.292	0	0	0	1	0	0	0	0	0	0	0	0	1
41	41	0.284	0	0	0	0	1	0	0	0	0	0	0	0	1
42	42	0.333	0	0	0	0	0	1	0	0	0	0	0	0	1
43	43	0.230	0	0	0	0	0	0	1	0	0	0	0	0	1
44	44	0.031	0	0	0	0	0	0	0	1	0	0	0	0	1
45	45	0.089	0	0	0	0	0	0	0	0	1	0	0	0	1
46	46	0.054	0	0	0	0	0	0	0	0	0	1	0	0	1
47	47	0.085	0	0	0	0	0	0	0	0	0	0	1	0	1
48	48	0.084	0	0	0	0	0	0	0	0	0	0	0	1	1
49	49	0.216	1	0	0	0	0	0	0	0	0	0	0	0	1
50	50	0.182	0	1	0	0	0	0	0	0	0	0	0	0	1
51	51	0.333	0	0	1	0	0	0	0	0	0	0	0	0	1
52	52	0.298	0	0	0	1	0	0	0	0	0	0	0	0	1

Before change

OUTPUT EXAMPLE TO SECTION IV D

OBS	TIME	CONC	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	C
53	53	0.272	0	0	0	0	1	0	0	0	0	0	0	0	1
54	54	0.193	0	0	0	0	0	1	0	0	0	0	0	0	1
55	55	0.174	0	0	0	0	0	0	1	0	0	0	0	0	1
56	56	0.158	0	0	0	0	0	0	0	1	0	0	0	0	1
57	57	0.096	0	0	0	0	0	0	0	0	1	0	0	0	1
58	58	0.056	0	0	0	0	0	0	0	0	0	1	0	0	1
59	59	0.117	0	0	0	0	0	0	0	0	0	0	1	0	1
60	60	0.201	0	0	0	0	0	0	0	0	0	0	0	1	1
61	61	0.283	1	0	0	0	0	0	0	0	0	0	0	0	1
62	62	0.320	0	1	0	0	0	0	0	0	0	0	0	0	1
63	63	0.254	0	0	1	0	0	0	0	0	0	0	0	0	1
64	64	0.272	0	0	0	1	0	0	0	0	0	0	0	0	1
65	65	0.145	0	0	0	0	1	0	0	0	0	0	0	0	1
66	66	0.189	0	0	0	0	0	1	0	0	0	0	0	0	1
67	67	0.095	0	0	0	0	0	0	1	0	0	0	0	0	1
68	68	0.075	0	0	0	0	0	0	0	1	0	0	0	0	1
69	69	0.052	0	0	0	0	0	0	0	0	1	0	0	0	1
70	70	0.046	0	0	0	0	0	0	0	0	0	1	0	0	1
71	71	0.071	0	0	0	0	0	0	0	0	0	0	1	0	1
72	72	0.010	0	0	0	0	0	0	0	0	0	0	0	1	1

OUTPUT EXAMPLE TO SECTION IV D

DEP VARIABLE: CONC

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	14	5.742797	0.410200	42.558	0.0001
ERROR	58	0.559039	0.009638597		
U TOTAL	72	6.301836			

ROOT MSE	0.098176	R-SQUARE	0.9113 ← R^2
DEP MEAN	0.240778	ADJ R-SQ	0.8914 ← R^2 adjusted for # of indipnt variables
C.V.	40.77468		

if less than or equal to α, shows that model is significant. For $\alpha = 0.05$, we conclude that model is significant

NOTE: NO INTERCEPT TERM IS USED. R-SQUARE IS REDEFINED.

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
X1	1	0.496496	0.044398	11.183	0.0001
X2	1	0.552580	0.044520	12.412	0.0001
X3	1	0.488331	0.044657	10.935	0.0001
X4	1	0.508082	0.044810	11.339	0.0001
X5	1	0.511333	0.044978	11.368	0.0001
X6	1	0.378917	0.045162	8.390	0.0001
X7	1	0.389222	0.046439	8.381	0.0001
X8	1	0.243806	0.046555	5.237	0.0001
X9	1	0.257057	0.046686	5.506	0.0001
X10	1	0.285308	0.046833	6.092	0.0001
X11	1	0.277559	0.046994	5.906	0.0001
X12	1	0.347477	0.047169	7.367	0.0001
C	1	-0.144327	0.040917	-3.527	0.0008
TIME	1	-0.0012509	0.0008483678	-1.474	0.1458

monthly effects — monthly means

all significantly $\neq 0$ at the 5% level

significant at the 5% level

not significant at the 5% level

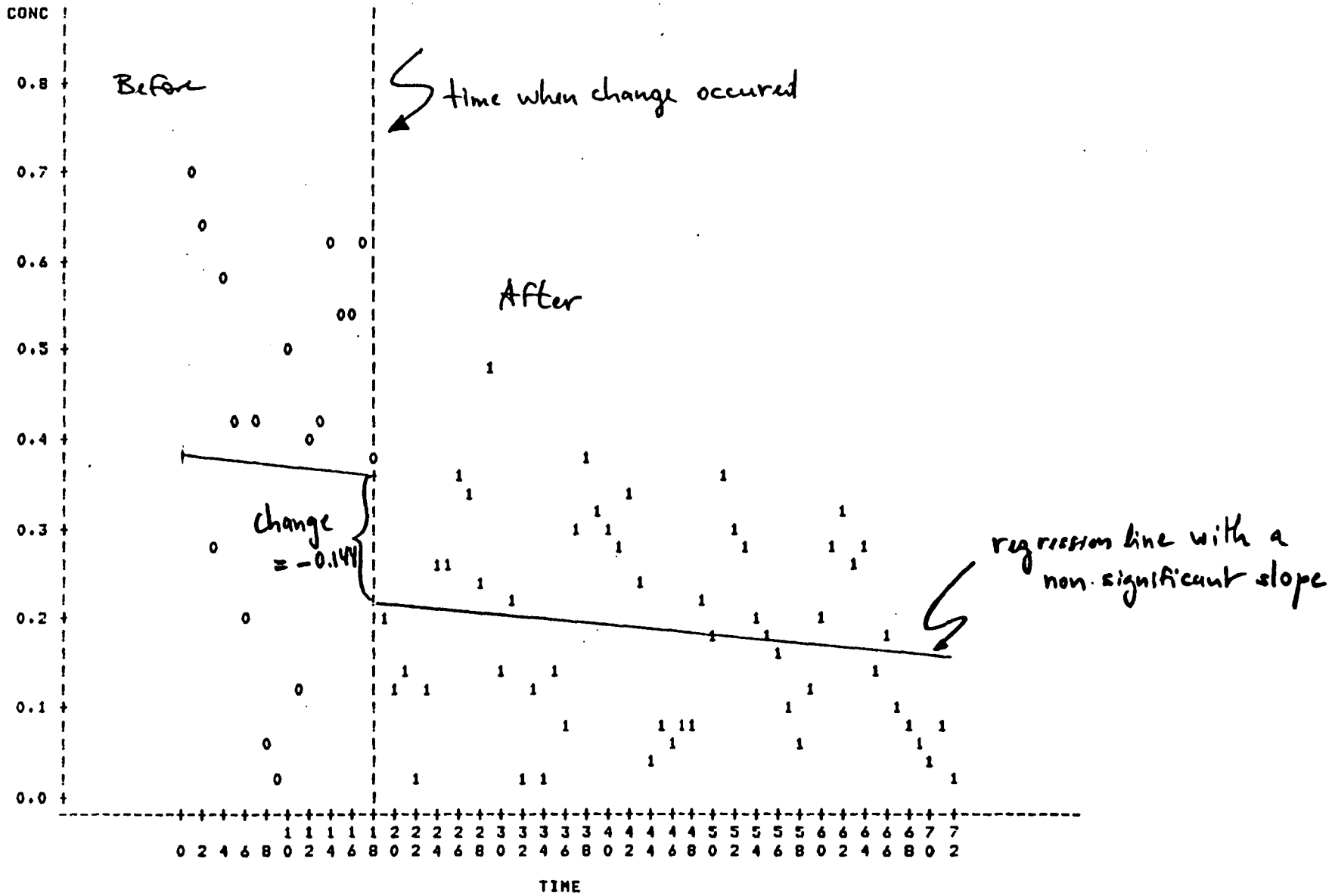
coefficient of change

coefficient of trend

MULTIPLE REGRESSION ANALYSIS 1

OUTPUT EXAMPLE TO SECTION IV D

PLOT OF CONC*TIME SYMBOL IS VALUE OF C



4. KENDALL'S TAU TEST (V-B)

```
OPTIONS LINESIZE=100 NODATE ;
DATA WATERQ;
INPUT MONTH WQI @@;
ORD=_N_;                                                          ①
CARDS ;
 1 21 2 3 3 5 4 8 5 21 6 48 7 37 8 39 9 26 10 16 11 35 12 7
TITLE KENDALL''S TAU TEST;
TITLE3 OUTPUT EXAMPLE TO SECTION V B;
PROC CORR DATA=WATERQ KENDALL;                                    ②
VAR MONTH WQI ;
*                                                                 ;
*THE FOLLOWING STEPS ARE TO COMPUTE THE N*(N-1)/2 SLOPES          ;
*THIS WILL ALLOW TO OBTAIN THE ESTIMATE OF THE TREND MAGNITUDE;
*IF THE TREND IS STATISTICALLY SIGNIFIGNANT                       ;
*                                                                 ;
DATA DIFFS ; SET WATERQ;
RETAIN X1-X50 -1
       Y1-Y50 -1;
ARRAY  XX (I) X1-X50;
ARRAY  YY (I) Y1-Y50;
    DO OVER XX;                                                   ③
       IF I=ORD THEN XX=MONTH;
       IF I=ORD THEN YY=WQI;
    END;
    DO OVER XX;
       IF I=ORD THEN GOTO LAB1;
       IF XX=-1 THEN GOTO LAB1;
       MONTH_A=XX;
       WQI_A=YY;
       OUTPUT;
    LAB1: END ;
DATA DIFFS ; SET DIFFS;
DROP X1-X50 Y1-Y50;
SLOPE=(WQI-WQI_A)/(MONTH-MONTH_A);
PROC RANK DATA=DIFFS ;                                            ④
VAR SLOPE ;
RANKS RK_SLOPE;
PROC SORT ; BY RK_SLOPE;
PROC PRINT ;                                                      ⑤
VAR MONTH_A MONTH WQI_A WQI SLOPE RK_SLOPE ;
PROC UNIVARIATE NOPRINT  DATA=DIFFS ;                             ⑥
VAR SLOPE;
OUTPUT OUT=MEDIAN
MEDIAN=MEDIAN;
PROC PRINT DATA=MEDIAN;
/*
```

① Data setup and data (If data is on file, use Infile statement)

② Correlation procedure with Kendall's tau option

③ Software to compute the $n(n-1)/2$ slopes with the given number, n, of pairs of data. Here $n=12$.

④ Ranking procedure and sorting procedure for the slopes computed in ③.

⑤ Optional; the output is not shown. If n is large, then the number of slops is large. With $n=12$, there are $12 \cdot 11/2 = 66$ slopes. With 24 months, there would be $24 \cdot 23/2 = 276$ slopes.

⑥ Univariate procedure with selected output to obtain the median slope value; with the print procedure.

KENDALL'S TAU TEST

OUTPUT EXAMPLE TO SECTION V B

sample size

VARIABLE	N	MEAN	STD DEV	MEDIAN	MINIMUM	MAXIMUM
MONTH	12	6.50000000	3.60555128	6.50000000	1.00000000	12.00000000
WQI	12	22.16666667	15.02623968	21.00000000	3.00000000	48.00000000

KENDALL TAU B CORRELATION COEFFICIENTS / PROB > |R| UNDER HO:RHO=0 / N = 12

	MONTH	WQI
MONTH	1.00000	0.19848
	0.0000	0.3716
WQI	0.19848	1.00000
	0.3716	0.0000

Kendall's tau

significance level of z, where z is the normal approximation of tau;
if greater than α, then tau is not significantly different from zero.

<u>Result</u>: Since 0.3716 is greater than 0.05 ($=\alpha$) we conclude that the trend (increasing by 1.49 units per year) is <u>not</u> significant

KENDALL'S TAU TEST

OUTPUT EXAMPLE TO SECTION V B

OBS	MEDIAN
1	1.48571

Note: This is printed on separate page

5. WILCOXON RANK SUM TEST (V-C)

```
OPTIONS LINESIZE=100 NODATE ;
DATA WATERQ;
INPUT TIME PERIOD $ CONC @@ ;
LABEL CONC=CONCENTRATION;
CARDS ;
1 B 99 2 B 111 3 B 74 4 B 123 5 B 71 6 B 75 7 B 59 8 B 85
9 A 59 10 A 99 11 A 82 12 A 51 13 A 48 14 A 39 15 A 42 16 A 42 17 A 47 18 A 50
PROC PLOT DATA=WATERQ;
PLOT CONC*TIME=PERIOD/HREF=9;
TITLE WILCOXON RANK SUM TEST (STEP TREND);
TITLE3 EXAMPLE OUTPUT TO SECTION V C;
PROC NPAR1WAY DATA=WATERQ  WILCOXON;
CLASS  PERIOD;
VAR CONC ;
/*
```

① Data set up

② Data

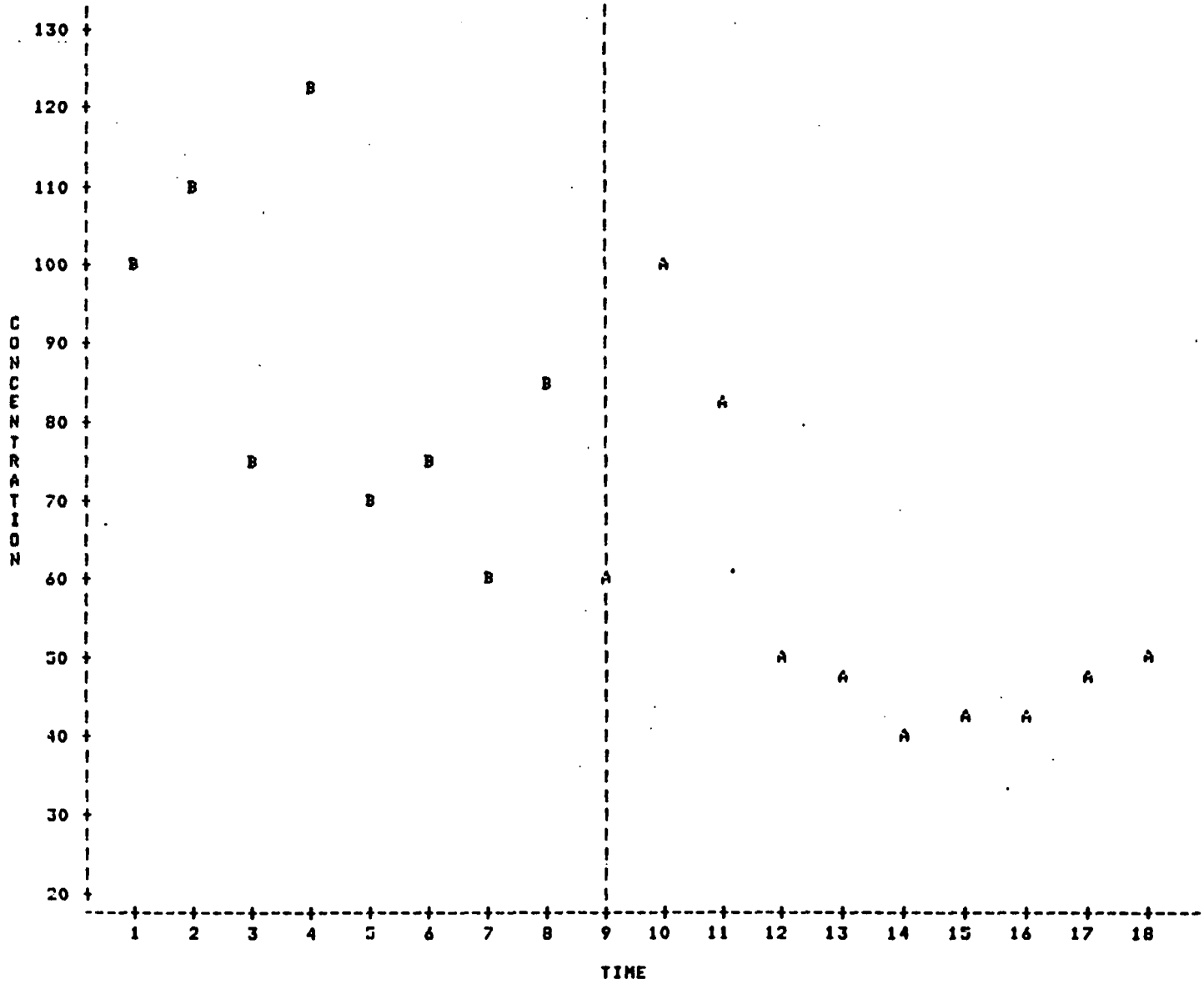
③ Plotting procedure

④ Procedure for computing the Wilcoxon statistic

WILCOXON RANK SUM TEST (STEP TREND) 1

EXAMPLE OUTPUT TO SECTION V C

PLOT OF CONC*TIME SYMBOL IS VALUE OF PERIOD



WILCOXON RANK SUM TEST (STEP TREND)

EXAMPLE OUTPUT TO SECTION V C

ANALYSIS FOR VARIABLE CONC CLASSIFIED BY VARIABLE PERIOD

AVERAGE SCORES WERE USED FOR TIES

WILCOXON SCORES (RANK SUMS)

LEVEL	N	SUM OF SCORES	EXPECTED UNDER HO	STD DEV UNDER HO	MEAN SCORE
B = Before	8	106.00	76.00	11.24	13.25
A = After	10	65.00	95.00	11.24	6.50

WILCOXON 2-SAMPLE TEST (NORMAL APPROXIMATION)
(WITH CONTINUITY CORRECTION OF .5)
S= 106.00 Z= 2.6252 PROB >|Z|=0.0087
T-TEST APPROX. SIGNIFICANCE=0.0177

KRUSKAL-WALLIS TEST (CHI-SQUARE APPROXIMATION)
CHISQ= 7.13 DF= 1 PROB > CHISQ=0.0076

Wilcoxon test statistic of 106 (= W in Section V-C)

significance (2-sided) of the normal transform of S

if less than or equal to α (e.g., 0.05) then reject the null hypothesis that slp trend is zero