Journal of Hazardous Materials, 18 (1988) 17-24
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

STATISTICAL ANALYSIS OF RARE EVENTS IN GROUNDWATER

CARL A. SILVER

Drexel University, Philadelphia, PA 19104 (U.S.A.) and Environmental Institute for Waste Management Studies, University of Alabama, Tuscaloosa, AL (U.S.A.)

and DENNIS DUNN

Drexel University, Philadelphia, PA 19104 (U.S.A.)

(Received December 6, 1986; accepted September 30, 1987)

Summary

Statistical analysis of trace organics and other pollutants that occur rarely in 'clean' groundwater poses difficult problems for conventional parametric statistical tests. The frequent occurrence of 'less than' values, values above the detection limit but below the quantification limit, many zeros, and unidentified pollutants make the use of statistics that require normal distribution or equal variances impractical. A method based upon the Poisson distribution that results in exact binomial probabilities for hypothesis testing is proposed. This method allows comparison of upgradient wells with down gradient wells or preoperational data with operational data for monitoring the performance of hazardous waste disposal sites. A numerical example, operating characteristic curves, and calculating algorithms are provided.

Introduction

Conventional approaches to statistical problems in groundwater monitoring, e.g., *t*-tests and analysis of variance, require normal distributions in the underlying variables. Although some non-normality can be tolerated under certain conditions, conventional tests fail badly when the distributions are severely truncated. Truncation occurs when data are cut off at an arbitrary level, as by a detection limit, or when a large proportion of values are zeros or 'not detected'. Often 'real' values, values above the detection limit, are but seldom encountered; that is, such values are rare events.

Rare events, as Mosteller and Rourke pointed out, have a known distribution. They wrote, "The distributions of the numbers of occurrences of events in fixed periods of time or space, or of counts of rare events, such as accidents, often conform approximately to a distribution called the Poisson" [1]. Under certain conditions occurrences of values of pollutants reliably above detection

0304-3894/88/$03.50

limits in groundwater monitoring at hazardous waste disposal sites follow the Poisson distribution.

Two types of testing are typical for monitoring groundwater near hazardous waste disposal sites. In one, monitoring depends upon comparison of water from wells upgradient of the waste disposal site with water from wells downgradient of the site. In the other, used where no clear gradient exists, water sampled from wells before a site has become operational (no on-site disposal has occurred) is compared with water sampled after disposal has begun. The statistical method described below is intended to be of use in both situations.

A chromatographic scan of water samples is often used to detect, identify, and if possible, quantify pollutants in groundwater. Typically, the chromatographic record is examined to determine whether any of a list of 'priority' pollutants is present. The chromatograhic record is compared with a library of known pollutants to determine if the record matches any 'signature' in the library. Usually, most of the 'priority' pollutants are not detected. Sometimes the scan indicates the presence of a pollutant but the chromatographic record does not match any known signature. Such a pollutant remains unidentified. Occasionally a record may indicate the presence of a pollutant at a value too low to be quantified. Other pollutants, identified or unidentified, may be present at a concentration sufficient for quantification. However, when the samples are from upgradient wells or from a preoperational time period there are usually few quantifiable values.

For pollutants that are unidentified it is impossible to test whether their concentrations have changed between upgradient and downgradient or between preoperational and operational time periods. Even for identified pollutants the occurrence of 'less than' values and markedly non-normal distributions make conventional parametric testing of means or magnitudes of concentrations inadvisable or impossible. We are proposing a method based upon 'events' where an event is defined as the occurrence of a pollutant above a preselected concentration. (See below for the treatment of values that are far above the preselected level.) The concentration selected may vary from pollutant to pollutant. In the proposed method one counts the number of 'events' and tests whether these events occur more often during the operational time period than they did preoperationally, or whether such events are more frequent downgradient than they are upgradient. If the events follow the Poisson distribution the crucial question is whether the parameter of the Poisson distribution has changed.

In its most general form, the Poisson distribution is applicable as a model for any counting process that comprises independent, rare events. Although the Poisson model is most often used for stochastic processes, it is not restricted to events in time. One form of the model, the "non-homogeneous Poisson process" [2] can be applied to the detection of groundwater constituents in a chromatographic scan.

Let E_i, $i=1,2,...$ be a set of dichotomous events and Let $P(E_i=1)=\upsilon_i$. The υ_i need not be assumed to be identical (hence the term "non-homogeneous"). In this application E_i registers the detection of a given constituent in a single sample of water. Let $X=E_1+E_2+...$, the total number of "detections" in a sample. Let $\lambda=\upsilon_1+\upsilon_2+...$, the expected number of "detections" in a given sample.

The number of detections, X, will be distributed according to the Poisson probability distribution,

$$P(X=k)=e^{-\lambda}\lambda^k/k!$$

if the following conditions are met:

1. The component events, E_i, are mutually independent.
2. The probability of occurrence of each event, υ_i, is arbitrarily small.
3. The total expected number of events $\upsilon_1+\upsilon_2+...$ is finite.

Condition 1 implies that the detection of one constituent in a well sample is unrelated to the detection of any other constituent. This assumption is certainly plausible if all detections result from random laboratory or measurement error, or if there is no necessary connection between the detection of one pollutant and the detection of some other one.

Conditions 2 and 3 imply that the events are rare. It is assumed that the detection of any given constituent is an unlikely event. In upgradient samples and in preoperational samples this assumption is well-founded. In any event, pollutants routinely found at a given site can be excluded from the analysis thus insuring that the conditions will be met. (Excluded pollutants can be statistically analyzed by conventional means.)

Two properties of the Poisson distribution yield a simple test statistic. If X_1, X_2, ..., X_n are independent Poisson variables having rate parameters λ_1, λ_2, ..., λ_n then the sum $Y=(X_1+X_2+...+X_n$ is Poisson distributed with rate parameter $\lambda=\lambda_1+\lambda_2+...+\lambda_n)$. Thus, the sum of all 'events' from n upgradient well samples (Y_1) is a Poisson variable. If all samples are assumed to be identically distributed with parameter λ, the overall rate is simply $n\lambda$. Similarly, under the null hypothesis, the total 'events' in a sample of m downgradient well samples (Y_2) is a Poisson variate with parameter $m\lambda$.

A second property of the Poisson distribution can be applied to the totals Y_1 and Y_2. If X_1 and X_2 are independent Poisson variables with parameters λ_1 and λ_2, then the conditional distribution of X_1, given X_1+X_2, is binomal with parameters $P=\lambda_1/(\lambda_1+\lambda_2)$ and $N=X_1+X_2$ [3] (see Appendix A). Under the null hypothesis that upgradient and downgradient samples come from the same population, P reduces to a simple function of the number of up and downgradient samples: $P=n/(n+m)$. A simple example based upon data from a "clean" site in the Northeast may help clarify the application of this property.

Table 1 shows the data upon which the following analyses are based. Two downgradient wells were not installed until the third quarter of 1986. There

TABLE 1

Number of times any compound was reported above Method Detection Limit (MDL)

Sampling quarter	Up gradient wells				Down gradient wells			
	U1	U2	U3		D2	D3	D4	D5
3rd 1985	1	3	1	0	1	1	NI	NI
4th 1985	0	0	0	0	0	0	NI	NI
1st 1986	2	0	0	0	0	0	NI	NI
2nd 1986	1	0	1	1	1	1	NI	NI
3rd 1986	1	1	0	0	0	0	0	1
4th 1986	1	1	0	0	1	1	1	1

NI = Not installed at this time.

are therefore 40 reports of samples (wells × quarters). A total of 23 events were reported. One should expect that because 18 of the 40 samples (45%) were from upgradient wells that $0.45 \times 23 = 10.35$ events would have occurred in upgradient samples and $0.55 \times 23 = 12.65$ of the 23 events would have occurred in downgradient samples. In fact, there were 13 upgradient hits and 10 events in downgradient samples. The probability of observing 13 or fewer events when one would expect 10.35, or 10 or more events when one would expect 12.65 is 0.9063 and can be obtained directly and exactly from a table of the binomial distribution with $P=0.45$, $Q=(1-P)=0.55$, and $N=23$. Table 2 shows the calculations required. Appendix B contains a Basic language program that will calculate the required probabilities for any values of P, Q, and N.

Alternatively, when all expected values are sufficiently large, e.g., five or more, the same probability can be easily approximated from the chi-square distribution (corrected for continuity) by calculating $\chi^2 = \Sigma(|O-E|+0.5)^2/E$ where O is the observed number of events in a time period and E is the expected number. In the present example this amounts to $(|13-10.35|+0.5)^2/10.35 + (|10-12.65|+0.5)^2/12.65 = 1.7431$. Reference to a table of chi-square with one degree of freedom shows that values of χ^2 that large or larger will occur approximately 90% of the time by chance when there is no difference between the upgradient and downgradient populations. One may also use the fact that when χ^2 has one degree of freedom $\chi^2 = Z^2$ where Z is a standard normal deviate. Thus, $Z = (1.7431)^{0.5} = 1.3203$. Reference to a table of the standard normal distribution shows the probability of obtaining 10 or more downgradient hits when 12.65 are expected by chance is 0.9066, in good agreement with the exact binomial probability of 0.9063 calculated above. These data, therefore, provide no evidence of an adverse environmental impact.

An important assumption of the method described herein is that the events are approximately Poisson distributed. The Kolmogorov–Smirnov D-Max statistic can be used to test that assumption [4]. An example of the calculation

TABLE 2

Calculations for the binomial distribution for 23 events: 13 upgradient and 10 downgradient

$N1$	$N2$	Factorials	coeff	prod	ind prob	cum $\geqslant$ N2
23	0	1	1	1.05654×10^{-8}	0.000000	1.000000
22	1	1	23	1.29133×10^{-8}	0.000000	1.000000
21	2	2	253	1.57829×10^{-8}	0.000004	1.000000
20	3	6	1771	1.92903×10^{-8}	0.000034	0.999996
19	4	24	8855	2.3577×10^{-8}	0.000209	0.999962
18	5	120	33649	2.88163×10^{-8}	0.000970	0.999753
17	6	720	100947	3.522×10^{-8}	0.003555	0.998783
16	7	5040	245157	4.30466×10^{-8}	0.010553	0.995228
15	8	40320	490314	5.26125×10^{-8}	0.025797	0.984675
14	9	362880	817190	6.43042×10^{-8}	0.052549	0.958878
13	10	3628800	1144066	7.8594×10^{-8}	0.089917	0.906329
12	11	39916800	1352078	9.60594×10^{-8}	0.129880	0.816412
11	12	479001600	1352078	1.17406×10^{-7}	0.158742	0.686533
10	13	6227020800	1144066	1.43496×10^{-7}	0.164169	0.527791
9	14	87178291200	817190	1.75384×10^{-7}	0.143322	0.363622
8	15	1307674368000	490314	2.14358×10^{-7}	0.105103	0.220300
7	16	20922789888000	245157	2.61994×10^{-7}	0.064230	0.115197
6	17	355687428096000	100947	3.20214×10^{-7}	0.032325	0.050967
5	18	6402373705728000	33649	3.91373×10^{-7}	0.013169	0.018642
4	19	121645100408830000	8855	4.78345×10^{-7}	0.004236	0.005473
3	20	2432902008176600000	1771	5.84644×10^{-7}	0.001035	0.001237
2	21	51090942171709000000	253	7.14565×10^{-7}	0.000181	0.000202
1	22	1124000727777600000000	23	8.73357×10^{-7}	0.000020	0.000021
0	23	25852016738885000000000	1	1.06744×10^{-6}	0.000001	0.000001

$N1$ = number of upgradient events; $N2$ = number of downgradient events; factorials = $N2!$; coeff = binomial coefficient = $23!/(N1!\times N2!)$; prod = $p^{N1}\times q^{N2}=0.45^{N1}\times0.55^{N2}$; ind prob = coeff × prod; and cum $\geqslant N2$ is the sum of ind prob from $N2$ to the total number of events.

of D-Max is given in Appendix C. Because there were 23 events in 40 samples we estimated λ to be $23/40=0.575$. The results of the calculations show the fit to the Poisson to be excellent. Tables for the interpretation of D-Max may be found in Refs. [4] or [5].

The procedure described in this paper is uniformly most powerful, i.e., no more powerful test is possible for the data and hypothesis considered above. The power of detecting a change in the rate of events depends upon the magnitude of the change and upon the total number of events. Figures 1 and 2 show the operating characteristics of the test when there are 23 of 100 total events respectively. In both figures π (the proportion of downgradient samples and, thus, the expected proportion of downgradient events under the hypothesis of no difference between up and downgradient populations) is set to 0.55.

The method described above is intended to be of use when detected events have an appreciable probability of being chance occurrences, i.e., random events

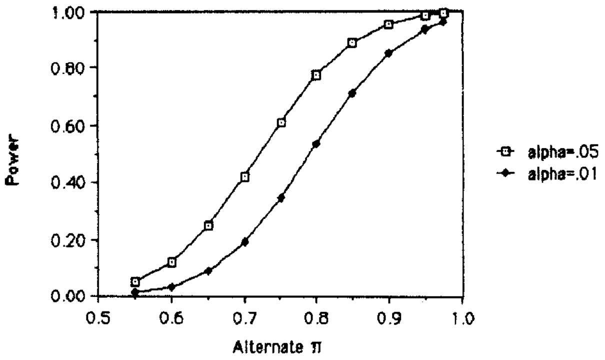

Fig. 1. Operating characteristic curve for binomial: $n = 23$, $\pi = 0.55$.

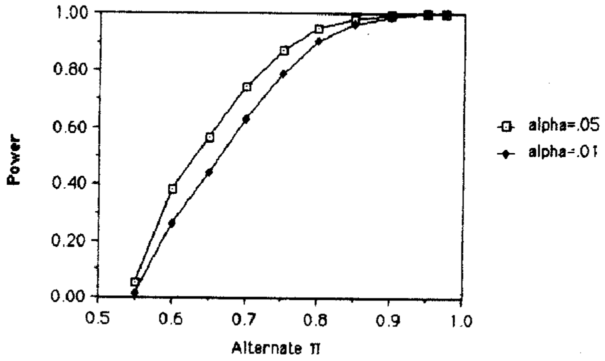

Fig. 2. Operating characteristic curve for binomial: $n = 100$, $\pi = 0.55$.

rather than pollution from a hazardous waste disposal site. Consequently it treats all events the same way; each is counted as a "1". To give appropriate weight to large concentrations that might indicate a real problem, it may be desirable to establish a concentration level above which a detection *per se* will trigger an appropriate action. A concentration that has a nearly vanishing probability of random occurrence should be selected for this purpose.

References

1 F. Mosteller, and R.E.K. Rourke, Sturdy Statistics, Addison-Wesley, Reading, MA, 1973.
2 E. Parzen, Stochastic Processes, Holden Day, San Francisco, 1962.
3 Y. Bishop, S. Feinberg and P. Holland, Discrete Multivariate Analysis, MIT Press, Cambridge, MA, 1975.
4 S. Siegel, Nonparametric Statistics, McGraw-Hill New York, NY, 1956.
5 Z.W. Birnbaum, Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample values, J. Amer. Stat. Assoc., 47 (1952) 425-441.

Appendix A

Let $Y_1 \sim \text{Poisson}(\lambda_1)$ and $Y_2 \sim \text{Poisson}(\lambda_2)$. Then the joint distribution of Y_1 and Y_2 is

$$f(Y_1,Y_2)=[(e^{-\lambda_1}\lambda_1{}^{Y_1})/Y_1!][(e^{-\lambda_2}\lambda_2{}^{Y_2})/Y_2!]$$

If we let $n=Y_1+Y_2$ the distribution of n is Poisson

$$f(n)=(e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n)/n!$$

and the conditional distribution of Y_1 and Y_2 given n is

$$f(Y_1,Y_2\,|\,n)=\frac{[(e^{-\lambda_1}\lambda_1{}^{Y_1})(e^{-\lambda_2}\lambda_2{}^{Y_2})]/Y_1!Y_2!}{(e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n)/n!}$$

$$=(n!/Y_1!Y_2!)[(\lambda_1/(\lambda_1+\lambda_2))^{Y_1}][(\lambda_2/(\lambda_1+\lambda_2))^{Y_2}]$$

which is the binomial distribution [5].

Appendix B

```
REM this program calculates binomial probabilities
REM N is the number of downgradient events; n is the total number of events.
PRINT "enter total number of events"
INPUT n
m=n+1
DIM fact(m), coeff(m), prod(m), prob(m), bum(m), cum(m)
PRINT "enter the proportion of downgradient reports"
PRINT "(equals the number of downgradient reports/total reports)"
INPUT p
q=1.00000-p
fact(0)=1
  FOR 1=1 TO n
    fact(i)=i*fact(i-1)
  NEXT i
    coeff(i)=fact(n)/(fact(i)*fact(n-i))
    prod(i)=(p↑(n-i))*(q↑(i))
    prob(i)=coeff(i)*prod(i)
  NEXT i
cum(0)=prob(0)
  FOR i=1 TO n
    cum(i)=prob(i)+cum(i-1)
  NEXT i
bum(0)=1.000000
  FOR i=1 TO n
    bum(i)=1-cum(i-1)
```

```
   NEXT i
LPRINT “Cumulative Probabilities of N or More Downgradient events [or”
LPRINT “Fewer than (n − N) Upgradient events] when the Probability of an”
LPRINT “Downgradient event is”;p
LPRINT
LPRINT
LPRINT
LPRINT “N”, “Ind. Prob.”, “Cum. Prob.”
   FOR i = 0 TO n
      LPRINT USING “##”;i,
      LPRINT “   ”,
      LPRINT USING “#######”;prob(i),
      LPRINT “   ”,
      LPRINT USING ”#######”;bum(i)
   NEXT i
END
```

Appendix C

Kolmogorov–Smirnov Worksheet

Events: 23; samples: 40; and λ: 0.575.

Poisson probabilities = $\exp(-\lambda) * (\lambda^{**}n)/n!$

n	Factorials	probability	exp cum prob	cum x	obs x	obs cum prob	diff
0	1	0.5627	0.5627	22.5082	20	0.5000	0.0627
1	1	0.3236	0.8863	35.4504	18	0.9500	0.0637
2	2	0.0930	0.9793	39.1713	1	0.9750	0.0043
3	6	0.0207	1.0000	40.0000	1	1.0000	0.0000

Table of D-Max critical values

p value	0.2	0.15	0.1	0.05	0.01
crit diff	0.1692	0.1802	0.1929	0.2150	0.2577

If maximum of diff excedes crit diff then reject Poisson