강 좌

수질 자료 분석을 위한 통계학적 방법

Statistical Methods for Water Quality Data Analysis

윤 광 식*
Yoon, Kwang Sik

1. 머 리 말

유역의 수질관리를 위해서는 수질 모니터링이 필수적이며, 얻어진 수질 자료는 자료분석 과정을 거치지 않으면 의미 없는 숫자이다. 유역의 여건을 고려한 자료해석이 필요하며 주관적인 해석을 피하기 위해서는 통계 기법의 활용이 필수적이다. 아무리 현장 실험이 사려깊게 계획되고 수행되더라도 측정한 자료는 불완전하고 불충분하게 된다. 불완전한 이유는 거의 대부분의 경우에 중요한 영향 변수들을 모두 알 수는 없기 때문이다.

올바른 통계치는 우리가 알려고 하지만 직접 관측할 수 없는 값을 추정하는데 이용될 수 있다. 통계를 이용하여 사실에 가깝게 다가갈 수는 있지만, 그것으로 사실에 도달될 것인지를 보장할 수는 없고, 또한 도달되었는지를 알 수도 없다. 그러나 통계를 이용하여, 어떤 가설이 참일 가능성에 대하여 과학적으로 설득력 있는 주장을 할 수 있게 된다.

사실과 과학의 추론 사이의 관계는 재판에서 무죄와 유죄가 아님의 판결과 유사한 면이 많다. 유죄가 아니라는 판결이 무죄가 증명되었음을 의미하는 것은 아니다. 그것은 단지 유죄가 증명되지 않았음을 의미할 뿐이다. 같은 논리로, 가설이 참일 가능성을 거부할 수 있는 지를 검정할 수 있을 뿐이다. 만약, 자료로부터 드러나는 사실에서 가설이 정말 같이 여겨지면, 우리는 가설이 참일 가능성에 기초를 두고 의사 결정을 해야한다. 또한, 진실이지만 증명되지 않은 가설을 거짓으로 잘못 판단하는데 따른 결과도 따져봐야 한다. 만약 그 결과가 심각할 수 있으면, 과학적 사실이 입증되지 않더라도 조치가 시행되어야 할 것이다.

수질 자료 분석 통계기법 활용 예로서, 1989년 위스컨신주의 수질에 관한 법규는 구체적으로 기하평균, 순위, 누적확률, 제곱합, 최소제곱회귀, 자료변환, 기하평균의 정규화, 결정계수, 유의수준 0.05에서 표준 F검정, 대표자료, 산술평균, 99백분위수, 확률분포, 대수정규분포, 계열상관, 평균, 분산, 표준절차, 표준정규분포, Z-값 같은 통계적 용어를 사용하였다. 실험자료의 통계적 분석에 관한 미국 환경보호청(US-EPA)의 지침문서에서는 아크사인변환, 프로빗분석, 비정규분포, Shapiro-Wilks 검정, 균질분산 Bartlett 검정, 이질분산, 반복(Replicate), Bonferroni 수정, t-검정, Dunnett 검정, Steel 순위검정, Wilcoxon 순위합검정 등을 언급하였다. 자연보호와 복원에 관한 법률(RCRA)의 대상지역에서 지하수 감시에 관한 미국의 EPA 지침문서에 언급된 용어에는 분산분석, 허용한계 단위, 예측구간, 관리도,

* 전남대학교 농과대학(농업과학기술연구소)

신뢰구간, Cohen 수정, 비모수 분산분석, 비율검정, α-오차, 검정력 곡선, 계열상관 등이 포함되어 있다. 이로부터, 수질 및 환경문제 규명을 위해 많은 통계기법이 이용되고 있음을 살펴볼 수 있다. 본 소고에서는 미국 농무성 산림국(USDA Forest service)에서 펴낸 『Statistical Methods Commonly Used in Water Quality Data Analysis』라는 책자의 예제를 통하여 수질 자료 해석을 위한 기본적인 통계처리 방법을 다루고자 한다.

2. 수질자료 통계 분석 기초

가. 수질 자료수가 다를 경우 두 모집단 평균비교(Comparison of Means from Two Populations when the Variance is Unknown and the Data are Unpaired) :

예제 (1) : 유역내 두 지천의 기저 유출시 부유유사 농도를 측정한 자료가 다음과 같다. 두지점의 측정 시기와 횟수가 다를 때 두 지점의 평균 농도값은 차이가 있는지 밝혀라.

Station A : 66, 59, 74, 60, 62, 69, 78, 71, 52, 78, 44, 50, 64

Station B : 61, 69, 67, 63, 39, 80, 63, 78, 47, 67, 72, 80, 41, 52, 64, 66, 74, 65, 67, 62

① 귀무가설 H_o 와 대립가설 H_a 를 설정

$H_o : \mu_A = \mu_B$

$H_a : \mu_A \neq \mu_B$

② 유의수준 설정 $\alpha = 0.05$

③ t-test를 위한 통계치 계산

$\overline{X}_A = 63.62$

$S_A = 10.62$

$\overline{X}_B = 63.85$

$S_B = 11.54$

$\overline{d} = \overline{X}_A - \overline{X}_B$

$\overline{d} = 63.62 - 63.85 = -0.23$

가중치 분산을 구하면

$$S_w^2 = \frac{S_A^2(n_A - 1) + S_B^2(n_B - 1)}{n_A + n_B - 2}$$

$$S_w^2 = \frac{(10.62)^2(12) + (11.54)^2(19)}{31}$$

$$S_w^2 = 125.3$$

$$S_{\overline{d}} = \sqrt{S_w^2\left(\frac{n_1 + n_2}{n_1 n_2}\right)}$$

$$= \sqrt{125.3\left(\frac{13 + 20}{13 \times 20}\right)}$$

$$= 3.998$$

$$t_s = \frac{\overline{d}}{S_{\overline{d}}}$$

$$t_s = \frac{-0.23}{3.988} = -0.058$$

④ 한계영역을 정의하면

$$t_c = \pm t_{1/2\alpha(n_1 + n_2 - 2)}$$

$t_c < t_{0.025(31)}$ 과 $t_c > t_{(1 - 0.025)(31)}$

그러므로 기각 영역은

$t < -2.042$ 와 $t > 2.042$

⑤ 계산된 t_s 값이 -2.042와 2.042 사이에 있으므로 귀무가설을 기각하지 않는다.

결론적으로 두측점의 부유유사의 평균값은 5% 유의수준에서 차이가 없다고 할 수 있다.

나. 수질 자료가 같을 경우 두 모집단 평균 비교 (Comparison of Means from Two Populations when the Variance is Unknown and the Data are Paired)

예제 (2) : 유역내 개발지역 상하류에서 용존산소 측정 결과가 다음과 같다. 유의수준 5%에서 상하류에서 수질 차이가 있는지 검토하라.

〈표 - 1〉 A · B의 용존산소 농도

용존산소 농도(mg/l)		차 이
Station A	Station B	d = A - B
6.2	5.2	1.0
6.5	5.4	1.1
6.8	5.3	1.5
7.0	5.7	1.3
6.9	5.6	1.3
7.0	6.2	0.8
6.8	5.7	1.1
6.7	5.6	1.1
6.8	5.8	1.0
6.2	5.6	0.6

① H_o 와 H_a를 설정

$H_o : \mu_A = \mu_B$

$H_a : \mu_A \neq \mu_B$

② 유의 수준 $\alpha = 0.05$

③ 통계량 t_s를 계산하면

$$\bar{d} = \bar{X}_A - \bar{X}_B$$

$$t_s = \frac{\bar{d}}{S_{\bar{d}} / \sqrt{n}}$$

$$t_s = \frac{1.008}{0.257 / \sqrt{10}}$$

$$t_s = 12.40$$

④ 한계영역을 정의하면

$$t_c = \pm t_{(\alpha/2)(n-1)}$$

$t_c < t_{0.025(9)}$ 과 $t_c > t_{0.975(9)}$

그러므로 기각영역은

$t < -2.262$ 와 $t > 2.262$

⑤ t_s 값이 12.4로 2.262 보다 크므로 귀무가설을 기각한다. T-test 결과 두 지천 상하류의 용존 산소의 평균 값은 유의수준 5%에서 다르다는 결론을 얻는다.

다. 자료수가 같은 경우 일원 배치 분산 분석 (One-Way Classification ANOVA with Equal Sample Size) :

예제 (3) : 유역내 산림 지역 2개소(측점 1과 2), 소 방목지(측점 3) 와 주택개발지(측점 4) 에서 질산태 질소 농도 측정 결과가 다음표와 같다. 유의수준 5%에서 측점별 질산태 질소 농도 평균값이 차이가 있는지 검토하라.

① 가설의 설정

$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_a : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

② 유의수준 결정, $\alpha = 0.05$

③ ANOVA 표를 작성하고 통계치를 결정
$F_S = 54.133$

〈표-2〉 측점별 질산태 질소 농도 (mg/ℓ)

측정	1	2	3	4
	8.7	9.8	11.9	8.8
	8.0	8.6	15.1	7.9
	8.9	8.8	11.2	8.5
	8.0	8.3	11.9	8.1
	6.8	7.4	13.9	10.0
	6.4	9.4	13.7	7.6
	7.8	7.9	12.6	10.1
	8.4	8.9	16.3	9.2
	7.8	8.3	15.4	10.0
	7.7	11.1	14.4	8.5
	8.3	8.9	13.2	12.7
	8.3	8.2	11.8	9.6
	9.7	10.7	12.6	8.5
	6.9	7.2	12.1	10.2
	7.4	7.2	13.3	6.6
ΣX	119.10	130.70	199.40	136.30
$\overline{X}$	7.94	8.71	13.29	9.09
S	0.85	1.16	1.50	1.44

④ 한계영역 F_c을 정의한다. 집단 간 자유도 4 - 1 = 3, 집단내 자유도 4 x (15-1) = 56

$F_c = F_{0.05(3,56)} \simeq 2.76$

⑤ F_S 값이 F_C 보다 크므로 귀무가설을 기각한다. 즉, 4곳의 측정 지점의 질산태 평균값이 모두 같지는 않다는 결론을 얻는다. 상기 분석으로는 어느 측점간 평균값이 통계적으로 다른지 알 수 없다.

라. 자료수가 같은 경우 Log 변환 일원 배치 분산 분석 (One-way Classification ANOVA with Equal Sample Size and Data Transformed by Log(X))

예제 (4) : 수질 영향에 의해 애벌레 숫자에 미치는 영향을 파악하기 위한 조사가 이루어졌다. 지천을 따라 수질 측정이 3개 지점에서 이루어졌으며, 측점 2와 측점 3 사이에는 마을 오수처리 연못에서 유입수가 있다. 조사된 각 측점의 개체(애벌레)수 평균간 차이가 있는지 검토하라.

분산이 실험 측정 값에 따라 서로 비슷한 범위에서 같지 않고 변하는 경우 분산이 비고정적이거나 불안정하다고 한다. 특히 미생물, 박테리아, 플랑크톤 개체수를 집계하는 경우 많이 발생한다. 이 경우 Log나 제곱근 변환을 통해 이를 교정한다.

〈표-3〉 측점별 애벌레 개체수

측점	1		2		3	
	개체수	Log (개체수)	개체수	Log (개체수)	개체수	Log (개체수)
	91	1.96	120	2.08	8	0.90
	77	1.89	110	2.04	17	1.23
	86	1.93	93	1.97	20	1.30
	52	1.72	150	2.18	15	1.18
	80	1.90	82	1.91	10	1.00
ΣX	386	9.40	555	10.18	70	5.61
$\overline{X}$	77.2	1.88	111.00	2.04	14.00	1.12
S	15.09	0.09	26.31	0.10	4.95	0.17

① 가설의 설정

$H_o : \mu_1 = \mu_2 = \mu_3$

$H_a : \mu_1 \neq \mu_2 \neq \mu_3$

② 유의수준 결정, $\alpha = 0.05$

③ ANOVA 표를 작성하고 통계치를 결정

이 문제의 해결을 위해서는 Log로 변환 값을 이용한다. $F_S = 75.97$

④ 한계영역 F_c을 정의한다. 집단 간 자유도 3 - 1 = 2, 집단내 자유도 3 x (5 - 1) = 12

$$F_c = F_{0.05(2,\ 12)} = 3.88$$

⑤ Fs > Fc 이므로 5% 유의 수준에서 귀무가설을 기각한다. 즉, 3 지점의 애벌레 측정수의 평균값은 모두 같지는 않다는 결론을 도출한다. 따라서, 오수에 의한 애벌레 수의 감소가 있는 것으로 추정할 수 있다.

마. 자료수가 다른 경우 Log 변환 일원 배치 분산 분석 (One-Way Classification ANOVA with Unequal Sample Size and Data Transformed by Log(X))

예제 (5) : 부유유사 농도를 산림(No. 1), 나지(No.2), 목초지(No. 3)의 세 군데 측정지점에서 측정하였다. 세 지점의 측정 횟수가 다를 때 토지이용에 따른 부유유사 농도의 평균값이 차이가 나는지 검토하라. (표본의 분산이 평균값과 독립적이지 않기에 Log 변환을 통해 이를 교정한다.)

① 가설의 설정

$H_o : \mu_1 = \mu_2 = \mu_3$

H_a : 적어도 한 지점의 평균값이 다른 지점의 값과 같지 않다.

② 유의 수준 결정, $\alpha = 0.10$

③ ANOVA 표를 작성하고 통계치를 결정

이 문제의 해결을 위해서는 Log로 변환 값을 이용한다. $F_S = 3.95$

〈표 - 4〉 측점별 부유 유사 농도

측점	1			2			3	
	mg/l	log(X)		mg/l	log(X)		mg/l	log(X)
	27	1.43		49	1.69		17	1.23
	29	1.46		25	1.40		30	1.48
	24	1.38		13	1.11		25	1.40
	48	1.68		29	1.46		18	1.26
	69	1.84		46	1.66		23	1.36
	30	1.48		15	1.18		18	1.26
	21	1.32		29	1.46		17	1.23
	68	1.83		23	1.36		10	1.00
	21	1.32		15	1.18		20	1.30
	20	1.30		28	1.45		21	1.32
	30	1.48	ΣX	272	13.95		37	1.57
	74	1.87	$\overline{X}$	27.2	1.40	ΣX	236	14.40
	26	1.41	S	12.3	0.20	$\overline{X}$	21.4	1.31
ΣX	487	19.80				S	7.3	0.15
$\overline{X}$	37.5	1.52						
S	20.1	0.21						

④ 한계영역 F_c을 정의한다. 집단 간 자유도 3 - 1 = 2, 집단내 자유도 34 - 3 = 31

$$F_c = F_{0.10(2,\ 31)} = 2.49$$

⑤ Fs > Fc 이므로 10% 유의 수준에서 귀무가설을 기각한다. 즉, 3 지점의 부유유사의 평균 농도값은 모두 같지는 않다는 결론을 도출한다

바. 이원배치 분산분석 (Two-Level Nested ANOVA)

예제 (6) : 질산태 질소 농도를 캠프장 오수가 유입되는 지점 상하류에서 측정하여 오수 유입 영향을 평가하고자 한다. 수질 분석 실험의 오차를 제거하기 위해 한 시료에 대해

3번 반복으로 수질분석을 실시하였다. 측점 간 수질 차이가 있는지 또 측점 내 반복수 간에 수질 차이가 있는지 검토하라.

① 가설의 설정

측점간 수질 비교

$H_o : \mu_A = \mu_B$

$H_a : \mu_A \neq \mu_B$

〈표-5〉 측점 AB의 질산태 질소 농도 (mg/ℓ)

측점 A.			측점 B.		
반복수			반복수		
1	2	3	1	2	3
1.0	1.1	1.1	5.1	5.3	5.1
1.6	1.5	1.6	6.0	5.8	6.1
1.3	1.3	1.3	5.8	5.9	5.9
1.4	1.3	1.3	6.5	6.5	6.4
1.5	1.5	1.6	6.7	6.6	6.8
2.0	1.8	1.9	6.1	6.1	6.1
2.1	2.0	2.0	6.9	6.8	6.9
1.7	1.8	1.7	5.5	5.5	5.6
1.6	1.6	1.6	5.4	5.4	5.4

실험 반복수 오차 검토

$H_o : \mu_1 = \mu_2 = \mu_3$

$H_a : \mu_1 \neq \mu_2 \neq \mu_3$

② 유의 수준 결정, $\alpha = 0.05$

③ ANOVA 표를 작성하고 통계치를 결정

측점간

$$F_s(site) = \frac{266.66667}{0.22181} = 1202.25$$

실험 반복수

$$F_s(num) = \frac{0.00370}{0.22181} = 0.02$$

④ 한계영역 F_c을 정의한다.

측점간 : 집단 간 자유도 2 - 1 = 1, 집단내 소집단간 자유도 2 x (3 - 1) = 4

$$F_c(site) = F_{0.05(1,4)} = 7.71$$

실험 반복수 : 집단내 소집단간 자유도 2 x (3 - 1) = 4, 소집단내 자유도 2 x 3 x (9 - 1) = 48

$$F_c(num) = F_{0.05(4,48)} = 2.56$$

⑤ $F_s(site) > F_c(site)$ 이므로 측점간 질산태 농도 평균이 같다는 귀무가설을 기각한다. 즉, 하류에 오수 유입 영향으로 질산태 농도가 증가한 것으로 본다.

$F_s(num) < F_c(num)$ 이므로 귀무 가설을 기각하지 않는다. 즉, 각 측정 샘플 반복간 평균 농도에는 차이가 없다는 결론을 얻는다.

〈표-6〉 부유유사농도 (mg/ℓ)

	Watershed	
Season	A	B
Spring - Summer	60 75 83 69 58 89	27 22 25 24 26 29
Fall - Winter	57 45 59 61 38 40	20 21 17 15 18 19

사. 반복이 있는 이원배치 분산분석 (Two-Way Classification ANOVA)

예제 (7) : 두 소유역의 출구에서 부유유사 농도를 1년간 측정하였다. 유역 A는 30%의 나지를 포함하고 있으며, 유역 B는 피복이 잘 되어 있는 유역이다. 두 소유역의 부유유사 평균 농도가 연간 또는 계절 별로 차이가 있는지 검토하라.

① 가설의 설정

측점 간 (Stations)

$H_o : \mu_A = \mu_B$

$H_a : \mu_A \neq \mu_B$

계절 간 (Seasons)

$H_o : \mu_{SS} = \mu_{FW}$

$H_a : \mu_{SS} \neq \mu_{FW}$

② 유의 수준 결정, $\alpha = 0.01$

③ ANOVA 표를 작성하고 통계치를 결정

측점간 F= 137.944

계절별 F= 19.481

측점과 계절 교호작용 F= 5.149

④ 한계영역 F_c을 정의한다.

측점, 계절, 교호 의 자유도가 모두 1 이므로

$F_c = F_{0.01(1, 20)} = 8.10$

⑤ 계절과 소유역의 경우 모두 $F_s > F_c$ 이므로 귀무가설을 기각한다. 따라서 유역간 그리고 계절 별로 부유유사 농도가 차이가 남을 알 수 있다. 유역과 계절의 교호관계는 유의성이 없다.

참고문헌

1. Averett, R. 1979. The use of select parametric statistical methods for the analysis of water quality data. Presented at USGS-BLM Conference on Water-Quality in Energy Areas. January 10~11, Denver, Colorado. p.16.
2. Calquhoun, D. 1971. Lectures on biostatistics. Claredon Press.
3. Elliott, J. M. 1971. Some methods for the statistical analysis of Benthic Invertebrates. Fresh Water Biol. Assoc. Sci. Pub. 25. p.144.
4. Freese, F. 1967. Elementary statistical methods for foresters. Agriculture Handbook 317. USDA - Forest Service. p.87.
5. Glass, G. V., P. D. Peckham and J. R. Sanders, 1972. Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Rev. Educ. Res. 42 : pp.237~288.
6. Ingwersen, J. B. 1981a. Statistical analysis using SAS at the USEPA National Computer Center. WSDG Application Document WSDG-AD-00001, USDA Forest Service, p.51.
7. Ingwersen, J. B. 1981b. Statistical analysis using SPSS at the USDA-Fort Collins Computer Center. WSDG Application Document WSDG-AD-00002, USDA Forest Service, p.9.
8. Nash, A. J. 1965. Statistical techniques in forestry. Lucas Brothers Publishers. Colu-

mbia, Missouri. p.146 Statistical Analysis System. 1979.

9. SAS Users Guide. SAS Institute Inc., 1980. P.O. Box 10066, Raleigh, North Carolina 17605. p.494.
10. Shapiro, S. S. and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). Biometrika. Vol. 52, pp. 591~611.
11. Sokal, R. R. and F. J. Rohlf. 1969. Biometry. W.H. Freeman and Company, San Francisco, p.776.
12. Statistical Package for the Social Sciencs, 1975. SPSS: Statistical Package for the Social Sciences. McGraw-Hill, Inc. New York, NY p.675.
13. Stephans, M. A. 1974. Use of the kolmogorov-Smirnov, Cramer-Von Mises and related statistics without extensive tables. J. American Statistical Association, 69 : p.730.
14. Steel, R. G .D. and J. H. Torrie. 1960. Principles and procedures of statistics with special reference to the biological sciences. McGraw-Hill Book Company, Inc. p.481.