기술강좌

수문통계학의 기초(Ⅲ)

허 준 행*

지난번 강좌에서는 확률현상, 임의사상, 확률변수, 확률밀도함수와 누가분포함수와 모집단과 표본에 대한 모멘트, 확률가중모멘트, L-모멘트의 정의 및 여러 가지 통계치 등 통계학의 기초에 대해 간단하게 설명하였으며, 연속되는 이번 강좌에서는 강우량 및 홍수량 같은 자료의 빈도해석에 대한 내용을 다루기로 한다. 이를 위해 경험적인 방법과 확률분포형을 이용한 방법에 대해 간단히 설명하기로 한다.

3. 검정방법

일반적으로 수문자료를 이용하여 통계적 분석을 할 때에는 대상 자료가 통계적 특성을 가지고 있는지 없는지 확인할 필요가 있다. 예를 들면, 홍수빈도 해석시 연홍수자료간에 무작위성(randomness)을 갖는지 확인하여야 하며, 대상자료가 정규분포여야 하는 경우에는 표본자료에 대하여 정규분포를 갖는지 확인하여야 한다. 또한 가정한 확률분포형을 이용하여 빈도해석을 실시하는 경우에는 대상자료가 가정한 확률분포형에 맞는지 확인하는 절차가 필요하게 된다. 따라서 본 강좌에서는 무작위성을 검정하는 4가지 방법, 대상 자료가 정규분포형인가를 확인하는 3가지 검정방법, 그리고 대상 자료가 가정 확률분포형에 적합한가를 판단하는 3가지 적합도 검정방법에 대해서 알아보기로 한다.

3.1 무작위성 검정(Randomness test)

자료의 무작위성을 검정하는 방법으로 본 절에서는 Anderson correlation test, Run test, Spearman's rank correlation coefficient test와 Turning point test에 대해서 간단히 설명하기로 한다.

3.1.1 Anderson correlation test

임의의 자료 $Y_1, \cdots Y_N$에서 자료수 N이 큰 경우에 자료의 자기상관계수 r_k는 평균값이 0이고 분산값이 1/N인 정규분포를 갖는다고 알려져 있다(Anderson, 1942). 그러므로 가설검정에서 귀무가설(null hypothesis) $H_0 : r_k = 0$, 대립가설(alternative hypothesis) $H_a : r_k \neq 0$라 설정하여 계산된 자기상관계수가 다음의 구간 안에 들어오는지 여부에 따라 검정할 수 있다.

$$\left[\frac{-u_{1-\alpha/2}}{\sqrt{N}}, \frac{u_{1-\alpha/2}}{\sqrt{N}}\right] \tag{3.1}$$

* 연세대학교 토목공학과 조교수

여기서 $u_{1-\alpha/2}$는 표준정규분포에서 $1-\alpha/2$ 분위수에 해당하는 값이다. 즉, 자료의 자기상관계수가 식 (3.1)의 범위안에 들어오면 대상자료는 α유의수준(α significance level)에서 무작위성(또는 독립성, independence)을 갖는다고 할 수 있는 것이다.

Anderson(1942)은 lag가 1인 상관계수의 평균과 분산을 다음 식과 같이 제안하였으며

$$E[r_1] = -\frac{1}{(N-1)} \tag{3.2}$$

$$Var(r_1) = \frac{(N-2)}{(N-1)^2} \tag{3.3}$$

Yevjevich (1972b)는 식 (3.2)와 (3.3)에서 N 대신에 N-k를 사용하여 lag가 k인 경우의 상관계수의 평균과 분산을 다음 식과 같이 수정하여

$$E[r_k] = -\frac{1}{(N-k)} \tag{3.4}$$

$$Var(r_k) = \frac{(N-k-1)}{(N-K)^2} \tag{3.5}$$

$\gamma = 1-\alpha$ 확률에 대한 구간을 다음과 같이 제안하였다.

$$\left[\frac{-1-u_{1-\alpha/2}\sqrt{N-k-1}}{N-k}, \frac{1-u_{1+\alpha/2}\sqrt{N-k-1}}{N-k}\right] \tag{3.6}$$

그러므로 귀무가설 $r_k=0$(대립가설 $r_k \neq 0$)에 대한 가설검정은, 상관계수 $r_k(k=1,\cdots, M)$의 값이 식 (3.6)의 구간을 벗어나는 총수가 $(1-\gamma)M$ 보다 크면 기각된다. 일반적으로 M의 값은 N/4를 이용한다.

【예제 3.1】 인도교지점의 연최대홍수량자료(1918-1991)는 표 1과 같다. 1952-1991년 사이의 연최대홍수자료가 자료 상호간에 독립성을 갖는지 알아보자.

표 1. 한강 인도교지점의 연최대홍수량(cms)

연 도	연최대홍수량(cms)									
1911								15997.6	15750.1	20813.6
1921	8949.8	19269.0	11931.2	16039.0	32971.8	17920.8	13692.5	4432.4	3932.8	17832.8
1931	10629.7	11720.3	11895.9	10106.1	21053.2	23074.9	10203.2	7289.8	668.6	22373.3
1941	—	—	—	—	—	—	—	—	—	—
1951	—	18874.6	16365.3	17517.6	13987.2	22338.0	12822.5	27409.4	24430.0	13750.2
1961	10438.4	13282.3	12791.1	11319.5	19399.8	19048.9	5310.1	9712.1	12791.1	11809.2
1971	6585.9	26431.3	2838.8	9997.5	12210.2	15922.2	3855.7	6819.1	14193.4	13200.2
1981	14811.6	5971.3	4420.4	28836.3	5883.8	6824.5	15531.4	9863.6	6083.1	32986.1
1991	14255.4									

먼저 인도교지점 연최대홍수량자료(1952-1991년)의 기본 통계값(unbiased statistics)은 아래와 같다.

SAMPLE MEAN=13772.980(cms)

STANDARD DEVIATION=7205.897

COEFFICIENT OF VARIATION=.5232

SKEWNESS COEFFICIENT=.7964

여기서 N=40이므로 M=10으로 정한다. 표 2에 lag가 10일 때까지 계산된 상관계수와 유의수준 $\alpha=0.05$일 때 식 (3.6)을 이용한 하한계와 상한계값을 나타내었다($u_{0.975}=1.96$). 표에서 보는 바와 같이 k=6일 때 r_6 이 상한계를 넘어가는 것을 볼 수 있다. $(1-\gamma)M=0.5$이므로 인도교지점의 홍수자료는 유의수준 $\alpha=0.05$에서 독립성을 갖지 못하지만 유의수준 $\alpha=0.01$에서는 독립성을 갖는다고 할 수 있다.

3.1.2 Run test

자료수가 N인 임의의 자료 $Y_1,\cdots Y_N$에 대해 식 (3.7)을 이용하여 0과 1로 정의되는 새로운 자료를 구할 수 있다.

표 2. 표본상관계수

Lag	하한계	r_k	상한계
1	-.310	-.113	.310
2	-.310	-.076	.310
3	-.310	.212	.310
4	-.310	.094	.310
5	-.310	-.136	.310
6	-.310	.311	.310
7	-.310	.053	.310
8	-.310	-.053	.310
9	-.310	.029	.310
10	-.310	.030	.310

ANDERSON TEST FOR INDEPENDENCE
TABULATED TEST VALUE=.500
COMPUTED TEST VALUE=1.000
SIGNIFICANCE LEVEL=.050
THE HYPOTHESIS OF INDEPENDENCE IS REJECTED

$$w_i = 1 \text{ if } Y_i > \bar{Y} \qquad (3.7a)$$
$$w_i = 1 \text{ if } Y_i < \bar{Y} \qquad (3.7a)$$

여기서 $\bar{Y}$는 자료의 평균값을 의미한다. 예를 들어, N=15인 경우 식 (3.7)에 의해 다음과 같은 임의의 값을 갖는 자료가 있다고 하자.

1 1 0 1 1 1 0 0 0 1 0 0 0 1 1

여기서 run은 연속적으로 0이거나 연속적으로 1인 경우를 말하며, 위 예에서 보면 0인 run은 3개이며, 1인 run은 4개로 총 7개의 run으로 구성되어 있다.

대상 자료가 무작위성(독립성)을 갖는 경우 전체 run수 U는 식 (3.8)의 평균과 식 (3.9)의 분산을 갖는 정규분포로 가정될 수 있다(Keeping, 1966).

$$E[U] = \frac{2N_1N_0}{N_1+N_0} + 1 \qquad (3.8)$$

$$Var(U) = \frac{2N_1N_0(2N_1N_0-N_1-N_0)}{(N_1+N_0)^2\,(N_1+N_0-1)} \qquad (3.9)$$

여기서 N_0와 N_1은 각각 0과 1의 개수이다. 이 경우 식 (3.10)의 검정통계량은 표준정규분포이다.

$$U_c = \frac{U-E[U]}{\sqrt{Var(U)}} \qquad (3.10)$$

그러므로 무작위성에 대한 귀무가설은 $|U_c| < u_{1-\alpha/2}$인 경우에 유의수준에서 받아들여진다.

【예제 3.2】 인도교지점의 연최대홍수량자료(1952-1991)에 대해 run test를 이용하여 연홍수자료간에 독립성을 갖는지 알아보자.

예제 3.1으로부터 평균값 $\bar{Y}$ =13772.98(cms)이므로, 식 (3.7)에서 정의된 w_i 자료는 아래와 같다.

0 0 0 0 0 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 0 1
1 1 0 1 1 0 1 0 1 1 0 1 1 0 1 1 0 0

여기서 N_0=17(0인 자료의 수), N_1=23(1인 자료의 수)이다. 또한 식 (3.8), (3.9), (3.10)에서 E[U]=20.55, Var(U)=9.29878, U_c=0.5083이고, 유의수준 α=0.05일 때 $u_{1-0.05/2}$=1.96이므로 $|U_x|\ u_{1-\alpha/2}$인 조건을 만족한다. 따라서 독립성에 대한 귀무가설은 유의수준 α=0.05에서 받아들여진다고 할 수 있다.

RUN TEST FOR INDEPENDENCE
TABULATED TEST VALUE=1.960
COMPUTED TEST VALUE=.508
SIGNIFICANCE LEVEL=.05
HYPOTHESIS OF INDEPENDENCE CANNOT BE REJECTED

3.1.3 Spearman's Rank Correlation Coefficient Test

Spearman의 순위상관계수 검정법은 자료수가 N인 원 자료 $Y_1, \cdots Y_N$의 자료 순서 $i(i=1,\cdots, N)$와 자료를 크기순으로(작은 값부터 큰 값 순으로) 재배열한 뒤, 크기순으로 배열하기전 원 자료의 순서 i에 해당하는 순서를 j로 정의하고, 하나의 자료에 대해 갖게 되는 두 개의 순위 (i, j)사이의 순위상관계수(rank correlation coefficient) R값에 기초하여 검정하는 방법으로 R은 다음 식과 같이 표현된다.

$$R=1-6\frac{\sum_{i=1}^{N}(i-j)^2}{N(N^2-1)} \quad (3.11)$$

만약 표본 자료가 독립성을 가지면, 식 (3.11)의 R은 표준정규분포이며, $1-R^2$은 자유도가 (N-2)인 χ^2-분포를 갖게 되므로 식 (3.12)의 통계량은 자유도가 (N-2)인 student t-분포를 갖게 된다.

$$T_c=\frac{R\sqrt{N-2}}{\sqrt{1-R^2}} \quad (3.12)$$

그러므로, 표본 자료의 독립성에 대한 귀무가설은 $|T_c| < t_{1-\alpha/2}(N-2)$ 인 경우 α유의수준에서 받아들여진다. 여기서 $t_{1-\alpha/2}(N-2)$는 student t-분포에서 자유도가 (N-2)일 때 $1-\alpha/2$분위수에 해당하는 값이다.

【예제 3.3】 인도교지점의 연최대홍수량자료에 대한 Spearman's rank correlation coefficient test 결과는 다음과 같다.

표 3에서 1952년부터 1991년 연최대홍수량자료의 순위(i)와 자료값은 ①, ②칸과 같으며 홍수자료를 크기순으로 재정열한 자료를 ④칸에 표기하였다. ③칸의 순위(j)는 순위홍수량이 크기순으로 정열되기 전의 순위(i)값에 해당된다. 그리고, ⑤칸의 값 (13859.0)은 식 (3.11)의 2번째 항 분자에 해당하는 값이므로 식 (3.11)의 R=-0.3001, 식 (3.12)의 $T_c=1.9393$이고 유의수준 $\alpha=0.05$에서 $t_{1-0.05/2}(38)=2.205$로 $|T_c| < t_{1-\alpha/2}(N-2)$ 인 조건을 만족하므로 독립성에 대한 귀무가설은 유의수준 $\alpha=0.05$에서 성립한다고 할 수 있다.

SPEARMAN'S COEFFICIENT TEST FOR INDEPENDENCE
TABULATED TEST VALUE=2.025
COMPUTED TEST VALUE=1.939
SIGNIFICANCE LEVEL=.050 HYPOTHESIS OF INDEPENDENCE CANNOT BE REJECTED

표 3. 인도교지점의 순위상관계수 계산예

순위(i) ①	홍수량 ②	순위(j) ③	순위홍수량 ④	$\Sigma(i-j)^2$ ⑤	순위(i) ①	홍수량 ②	순위(j) ③	순위홍수량 ④	$\Sigma(i-j)^2$ ⑤
1	18874.6	32	2838.8	961.0	21	26431.3	37	13200.2	7456.0
2	16365.3	30	3855.7	1745.0	22	2838.8	1	13282.3	7897.0
3	17517.6	31	4420.4	2529.0	23	9997.5	13	13750.2	7997.0
4	13987.2	24	5310.1	2929.0	24	12210.2	17	13987.2	8046.0
5	22338.0	35	5883.8	3829.0	25	15922.2	29	14193.4	8062.0
6	12822.5	20	5971.3	4025.0	26	3855.7	2	14255.4	8638.0
7	27409.4	38	6083.1	4986.0	27	6819.1	9	14811.6	8962.0
8	24430.0	36	6585.9	5770.0	28	14193.4	25	15531.4	8971.0
9	13750.2	23	6819.1	5966.0	29	13200.2	21	15922.2	9035.0
10	10438.4	14	6824.5	5982.0	30	14811.6	27	16365.3	9044.0
11	13282.3	22	9712.1	6103.0	31	5971.3	6	17517.6	9669.0
12	12791.1	18	9863.6	6139.0	32	4420.4	3	18874.6	10510.0
13	11319.5	15	9997.5	6143.0	33	28836.3	39	19048.9	10546.0
14	19399.8	34	10438.4	6543.0	34	5883.8	5	19399.8	11387.0
15	19048.9	33	11319.5	6867.0	35	6824.5	10	22338.0	12012.0
16	5310.1	4	11809.2	7011.0	36	15531.4	28	24430.0	12076.0
17	97.12.1	11	12210.2	7047.0	37	9863.6	12	26431.3	12701.0
18	12791.1	18	12791.1	7047.0	38	6083.1	7	27409.4	13662.0
19	11809.2	16	12791.1	7056.0	39	32986.1	40	28836.3	13663.0
20	6585.6	8	12822.5	7200.0	40	14255.4	26	32986.1	13859.0

3.1.4 Turning Point Test

자료수가 N인 임의의 자료 $Y_1, \cdots Y_N$에서 peak와 trough는 각각 식 (3.13)과 (3.14)와 같이 정의된다.

$$Y_{i-1} < Y_i > Y_{i+1} \qquad (3.13)$$
$$Y_{i-1} > Y_i < Y_{i+1} \qquad (3.14)$$

만약 표본자료가 무작위성을 갖는다면, peak와 trough의 총수 M은 각각 식 (3.15)와 (3.16)의 평균과 분산을 갖는 정규분포가 된다(Clarke, 1973).

$$E[M] = \frac{2(N-2)}{3} \qquad (3.15)$$
$$Var(M) = \frac{(16N-29)}{90} \qquad (3.16)$$

그러므로 식 (3.17)의 검정통계량은 표준정규변수가 된다.

$$U_c = \frac{M-E[M]}{\sqrt{Var(M)}} \qquad (3.17)$$

이 경우, 표본자료의 독립성에 대한 귀무가설은 $|U_c| < u_{1-\alpha/2}$인 경우 α유의수준에서 받아들여진다.

【예제 3.4】 인도교지점의 연최대홍수량자료에 대한 Turning point test 결과는 다음과 같다. 인도교지점의 홍수자료 중 식 (3.13)과 (3.14)를 만족하는 peak와 trough의 총수는 M=26개이고, 식 (3.15), (3.16)에서 M의 평균=25.33, M의 표준편차=2.61이므로 식 (3.17)에서 $U_c=0.256$이 된다. $u_{1-0.05/2}=1.960$이므로 $|U_c| < u_{1-\alpha/2}$인 조건을 만족하므로 자료의 독립성은 α유의수준에서 받아들여진다.

TURNING POINT TEST FOR INDEPENDENCE
TABULATED TEST VALUE=1.960
COMPUTED TEST VALUE=.256
SIGNIFICANCE LEVEL=.050 HYPOTHESIS OF INDEPENDENCE CANNOT BE REJECTED

3.2 정규분포 검정

본 절에서는 대상 표본자료가 정규분포를 갖는지 여부를 검정하는 방법에 대하여 설명하기로 한다.

3.2.1 왜곡도계수 검정(Skewness test of normality)

자료가 정규분포인 경우의 왜곡도계수는 0의 값을 갖는다. 자료수가 N인 임의의 자료 $Y_1, \cdots Y_N$의 불편의 왜곡도계수는 다음 식을 이용하여 구할 수 있다.

$$\bar{\gamma} = \frac{N\sum_{i=1}^{N}(Y_i-\bar{Y})^3}{(N-1)\,(N-2)s_Y^3} \qquad (3.18)$$

여기서 $\bar{Y}$는 평균값이고, s_Y는 표준편차이다. 표본자료의 수가 충분히 큰 경우(N>150) 왜곡도계수는 평균이 0, 분산이 6/N인 정규분포를 갖게 된다(Snedecor와 Cochran, 1980). 그러므로, 표본 왜곡도계수가 식 (3.19) 안에 들어오면 α유의수준에서 정규분포라고 할 수 있다.

$$\left[-u_{1-\alpha/2}\sqrt{\frac{6}{N}},\ u_{1-\alpha/2}\sqrt{\frac{6}{N}}\right] \qquad (3.19)$$

만약에 표본자료 수가 작은 경우에는 주의가 필요하며 일반적으로 Pearson과 Hartley(1966)이 제안한 표값 $g_{1-\alpha/2}(N)$을 사용한다. 여기서 $g_{1-\alpha/2}(N)$은 유의수준 α와 자료수 N의 함수로 표 4와 같다. 그러므로 $|\bar{\gamma}| > g_{1-\alpha/2}(N)$이면 정규분포에 대한 귀무가설은 기각된다.

【예제 3.5】 인도교지점의 연최대홍수량자료에 대한 왜곡도계수 검정 결과는 다음과 같다. 인도교지점 홍수자료의 불편의 왜곡도계수 $\bar{\gamma}=0.7964$이고, 유의수준 $\alpha=0.02$일 때 $g_{1-0.02/2}(40)=0.870$이므로 정규분포에 대한 귀무가설은 받아들여진다.

TABULATED TEST VALUE=.870
COMPUTED TEST VALUE=.796
SIGNIFICANCE LEVEL=.020
HYPOTHESIS OF NORMALITY CANNOT BE REJECTED

3.2.2 Shapiro-Wilk 검정

이 방법은 Shapiro와 Wilk (1965)에 의해서 소개된 검정 방법으로 정규분포검정에 좋다고 알려져 있으며, 초기에는 자료수가 50보다 작은 경우에만

표 4. Quantile Pointsg_β (N) for the Distribution of the Sample Skewness Coefficient(Pearson and Hartley, 1966)

Sample size	Cumulative Probability(β)		Sample size	Cumulative Probability(β)		Sample size	Cumulative Probability(β)	
N	0.95	0.99	N	0.95	0.99	N	0.95	0.99
25	.711	1.061	200	.280	.403	900	.134	.190
30	.662	.986	250	.251	.360	950	.130	.185
35	.621	.923	300	.230	.329	1000	.127	.180
40	.587	.870	350	.213	.305	1200	.116	.165
45	.558	.825	400	.200	.285	1400	.107	.152
50	.534	.787	450	.188	.269	1600	.100	.142
60	.492	.723	500	.179	.255	1800	.095	.134
70	.459	.673	550	.171	.243	2000	.090	.127
80	.432	.631	600	.163	.233	2500	.080	.114
90	.409	.596	650	.157	.224	3000	.073	.104
100	.389	.567	700	.151	.215	4000	.064	.090
125	.350	.508	750	.146	.208	5000	.057	.081
150	.321	.464	800	.142	.202			
175	.298	.430	850	.138	.196			

* Negative values of the quantile points correspond to lower limits

사용 가능하였지만 후에 Royston(1982)이 자료수가 2000인 경우까지 사용할 수 있도록 보완하였다. 자료수가 N인 임의의 자료 $X_1, \cdots X_N$의 Shapiro-Wilk 검정통계량은 식 (3.15)와 같이 정의된다.

$$W = \frac{\left[\sum_{i=1}^{k} a_i (Y_{N-i+1} - Y_i)\right]^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} \tag{3.20}$$

여기서 Y는 표본자료 X를 작은 값부터 크기 순으로 재정열하였을 때에 해당하는 순위통계값(order statistic)이고 a는 정규화된 최적선형불편의계수(best linear unbiased coefficient)이며(Shapiro와 Wilk, 1965), 자료수가 짝수인 경우 k=N/2이고 홀수인 경우에는 k=(N-1)/2이다. Monte Carlo모의발생을 통하여 Royston(1982)은 식 (3.21)에 정의된 변수가 정규분포임을 보였다.

$$Z = (1 - W^{\lambda}) \tag{3.21}$$

여기서 λ는 다항식으로 주어진다(Royston, 1982). 이와 마찬가지로 식 (3.22)에 주어진 변수도 정규분포를 갖게 된다.

$$U = \frac{Z - \bar{Z}}{\sigma_z} \tag{3.22}$$

여기서 $\bar{Z}$와 σ_z는 각각 변수 Z의 평균 및 표준편차이다. 그러므로, 표본자료로부터 W가 주어지면 식 (3.21)과 (3.22)를 이용하여 Shapiro-Wilk의 W 통계량의 유의수준을 구할 수 있다. 계산된 U 값이 큰 경우에(또는 계산된 유의수준 α값이 0.01 또는 0.05 보다 작은 경우) 대상 표본자료가 정규분포라는 귀무가설은 기각된다.

【예제 3.6】 인도교지점의 연최대홍수량자료에 대한 Shapiro Wilk test 결과는 다음과 같다. 식 (3.21)을 이용한 인도교지점 홍수자료의 W=0.938, λ=0.131, $\bar{Z}$ =0.628, σ_Z=0.038이므로 U=1.754이다. 이 값을 유의수준으로 환산하면 0.0397이 된다.

SHAPIRO-WILK TEST FOR NORMALITY
W TEST STATISTIC=.938
STANDARD NORMAL DEVIATE=1.754
SIGNIFICANCE LEVEL=.397E-01

그러므로, 인도교지점 홍수자료에 대한 Shapiro Wilk 검정방법 적용 결과 유의수준 $\alpha=0.01$일 때는 정규분포라고 할 수 있지만, $\alpha=0.05$에서는 정규분포에 대한 귀무가설은 기각된다.

3.2.3 Probability Plot Correlation Coefficient Test

Probability plot correlation coefficient (PPCC) 검정방법은 Filliben(1975)에 의해 처음 제안되어 Filliben 검정방법이라고도 불리며 간단하고 편리하면서도 대상 표본자료가 정규분포인가를 판단하는데 좋은 방법으로 알려져 있다(Vogel, 1986). 또한 정규분포 외에도 일반적으로 2개의 매개변수를 갖는 분포형에도 적용이 가능하다(Vogel, 1986).

자료수가 N인 자료를 $Y_1,\cdots Y_N$, 이에 상응하는 순위통계값을 $X_1,\cdots X_N$ 이라 하고 표준정규분포로부터 중앙값(median)을 M_i 라고 하자. 표본자료 Y 가 정규분포를 가지면, X_i와 M_i를 그림상에 표시하였을 때 거의 직선으로 나타나게 된다. 이때에 식 (3.23)으로 정의되는 적모멘트상관계수(product moment correlation coefficient)가 X_i와 M_i 사이의 선형성의 척도를 나타내게 된다.

$$\rho_c=\frac{\sum_{i=1}^{N}(X_i-\bar{X})\ (M_i-\bar{M}_i)}{\sqrt{\sum_{i=1}^{N}(X_i-X)^2\sum_{i=1}^{N}(M_i-\bar{M}_i)^2}} \tag{3.23}$$

식 (3.23)에서 M_i는 다음 식 (3.24)에 의해 구해지며

$$M_i=\Phi^{-1}(m_i) \tag{3.24}$$

m_i(uniform order statistic median)는 다음 식에 의해 주어진다.

$$\begin{aligned} m_i&=1-(0.5)^{1/N}\quad i=1\\ &\frac{(i-0.3175)}{(N+0.365)}\quad i=2,\cdots,N-1\\ &(0.5)^{1/N}\quad i=N\end{aligned} \tag{3.25}$$

여기서 Φ^{-1}은 표준정규누가함수의 역함수이다. PPCC검정방법의 귀무가설 $H_0:\rho_c=1$, 대립가설 $H_a:\rho_c<1$로 일방향 가설검정방법(one-side hypothesis test)을 사용한다. 그러므로 표본자료가 정규분포라는 가설은 식 (3.26)의 조건을 만족할 때 성립되며 이와 반대인 경우에는 정규분포가 아니라고 할 수 있다.

$$\rho_c>r_\alpha(N) \tag{3.26}$$

여기서 $r_\alpha(N)$ 은 자료수가 N인 경우 PPCC검정통계량을 나타내는 것으로 표값으로 주어지게 된다(Filliben, 1975; Vogel, 1986 참조).

【예제 3.7】 인도교지점의 연최대홍수량자료에 대한 PPCC test 결과는 다음과 같다. 식 (3.25), (3.24)에서 계산된 m_i와 m_i는 표 5와 같다. 인도교지점의 홍수자료 평균값 $\bar{X}=13772.98$(예제 3.1 참조)이고 m_i 값의 평균값 $\bar{M}=0.0$이므로, 식 (3.23)으로부터 계산된 $\rho_c=0.9718$이다. 여기서 $\rho_c=0.9718<r_{0.05}(40)=0.972$ 이므로 정규분포에 대한 가설은 기각된다. 그러나 만약에 유의수준을 $\alpha=0.01$ 로 하면, $r_{0.01}(40)=0.958$ 이기 때문에 $\rho_c>r_{0.05}(40)$ 이 되어 정규분포에 대한 귀무가설은 받아들여진다.

FILLIBEN TEST OF NORMALITY
TABULATED TEST VALUE=.97200
COMPUTED TEST VALUE=.97198
SIGNIFICANCE LEVEL=.05000
THE HYPOTHESIS OF NORMALITY IS REJECTED

3.3 적합도 검정(Goodness of Fit Test)

적합도 검정은 대상자료로부터 얻어지는 경험적 빈도분포와 가정한 확률분포가 얼마나 잘 일치하는

표 5. PPCC검정방법 계산예

No.	m_i	M_i	No.	m_i	M_i
1	.017	-2.116	21	.512	.031
2	.042	-1.732	22	.537	.093
3	.066	-1.503	23	.562	.156
4	.091	-1.333	24	.587	.219
5	.116	-1.195	25	.611	.283
6	.141	-1.077	26	.636	.348
7	.166	-.972	27	.661	.415
8	.190	-.877	28	.686	.484
9	.215	-.789	29	.711	.555
10	.240	-.706	30	.735	.629
11	.265	-.629	31	.760	.706
12	.289	-.555	32	.785	.789
13	.314	-.484	33	.810	.877
14	.339	-.415	34	.834	.972
15	.364	-.348	35	.859	1.077
16	.389	-.283	36	.884	1.195
17	.413	-.219	37	.909	1.333
18	.438	-.156	38	.934	1.503
19	.463	-.093	39	.958	1.732
20	.488	-.031	40	.983	2.116

가를 판단하는 방법으로 도식적인 방법과 통계적인 방법에 의해 이루어진다. 도식적인 방법의 대표적인 예는 확률지로, 예를 들어 대상 표본자료의 경험적 누가분포가 정규확률지상에서 직선으로 나타나면, 이 자료는 정규분포를 갖는다고 할 수 있다. 그러나 도식적인 방법은 주관적인 경우가 많으므로, 보다 객관적인 방법으로 통계적인 방법이 널리 사용되고 있다.

본 절에서는 통계적인 적합도 검정 방법으로 많이 사용하는 χ^2방법, Kolmogorov- Smirnov방법, Cramer von Mises방법에 대해서 설명하기로 한다.

3.3.1 χ^2-검정

χ^2-적합도 검정은 가장 널리 사용하는 적합도 검정방법 중의 하나로, 대상 자료에 대해 적합하다고 가정한 확률밀도함수와 전편(II)에서 설명한 군집화된 자료를 이용한 빈도해석을 통하여 구해지는 경험적 확률밀도함수를 비교하는 검정방법이다.

r개의 매개변수 $\underline{\theta}(\theta_1,\cdots,\theta_r)$를 갖는 확률분포의 확률밀도함수를 $f(y,\underline{\theta})$, 누가분포함수를 $F(y,\underline{\theta})$라 하자. χ^2-적합도 검정절차는 아래와 같다.

전체확률구간을 p_1 에서 p_m까지의 확률을 갖는 m개의 계급구간으로 나눈다($\sum_{i=1}^{m} p_i=1$). 임의의 계급구간 $j(j=1,\cdots,m)$에 들어가는 절대빈도수 N_j를 계산한다.

임의의 계급구간 j에 들어가는 기대자료수(expected number$=p_jN$)를 계산한다.

자유도가$(m-r-1)$이고 χ^2-분포를 갖는 검정통계량 C를 다음 식 (3.27)를 이용하여 계산한다.

$$C=\sum_{j=1}^{m}\frac{(N_j-p_jN)^2}{p_jN} \tag{3.27}$$

대상 표본자료에 대하여 $C>\chi^2_{1-\alpha/2}(m-r-1)$인 조건을 만족하면 유의수준 α에 대해 가정한 확률밀도함수가 대상 표본자료에 맞는다고 할 수 있다.

【예제 3.8】 인도교지점의 연최대홍수량자료에 대한 χ^2-test 결과는 다음과 같다.

정규분포가 인도교지점 홍수자료에 적합하다고 가정하여 모멘트법을 이용하여 매개변수를 추정하면 다음과 같다 (최우도법을 이용해도 동일한 매개변수를 가짐).

PARAMETERS OF THE NORMAL DISTRIBUTION (UNBIASED MOMENT AND MAX. LIKELIHOOD ESTIMATES)
LOCATION PARAMETER=13772.98000(MEAN)
SCALE PARAMETER=7205.89700(STD DEV.)

여기서 계급구간은 전편(II)의 식 (2.9)를 이용하면 $N_c=1+3.322\ \log_{10}(40)=6.32$ 이므로 6개 구간으로 정한다. 따라서 $p_1=p_2=\cdots=p_6=0.1667$이고 이에 대한 홍수량과 구간별 관측자료수 및 기대값은 표 6과 같다. 식 (3.27)을 이용하여 계산된 검정통계값은 C=2.30이고 자유도가 3(=6-2-1), 유의수준 $\alpha=0.05$일 때의 $\chi^2_{1-0.05/2}(3)=7.81$보다 작으므로 표본자료의 정규분포에 대한 귀

무가설은 성립한다.

표 6. χ^2-검정방법의 계산예

구 간 (j)	CDF	구간 j의 상한값	구간j에서의 관측자료수 (N_j)	구간j에서의 기대값 (p_i N)
1	.1667	6802.26	8	6.7
2	.3333	10672.34	6	6.7
3	.5000	13772.98	9	6.7
4	.6667	16873.62	7	6.7
5	.8333	20743.70	4	6.7
6	1.0000	∞	6	6.7

COMPUTED TEST VALUE=2.300
CHI-SQUARE TABLE VALUE=7.810
SIGNIFICANCE LEVEL=.050
HYPOTHESIS OF GOOD FIT CANNOT BE REJECTED

3.3.2 Kolmogorov-Smirnov 검정

Kolmogorov-Smirnov 검정방법은 앞에서 설명한 χ^2-검정방법과 달리 확률밀도함수 대신에 누가분포함수에 대해 검정하는 방법이다.

이 방법은 경험적 누가분포함수 $F_e(y)$ 와 가정한 누가분포함수 $F(y;\underline{\theta})$ 사이의 최대차이값에 의해 적합성을 검정하는 방법으로 검정통계량은 다음 식과 같이 주어진다.

$$D=\mathrm{Max}\mid F_e(y_i)-F(y;\underline{\theta})\mid \tag{3.28}$$

여기서 경험적 누가분포함수를 구하는 도시공식으로는 i/N이 사용되었으나 실제로는 i/(N+1)이 많이 쓰여지고 있다(Yevjevich, 1972a). Kolmogorov-Smirnov 검정방법을 통한 가설검정은 유의수준 α에서 $D<d_{1-\alpha/2}(N)$를 만족하면 받아들여진다. 여기서 $d_{1-\alpha}(N)$은 KolmogorovSmirnov통계량으로 표 7과 같다.

또한 Kolmogorov-Smirnov 검정방법은 두 종류의 표본자료가 동일한 확률분포를 갖는지 여부를 판단하는 데에도 사용될 수 있다. 이 경우의 검정통계량은 다음과 같이 주어진다(Benjamin과 Cornell, 1970).

$$D^*=\mathrm{Max}\mid F_{e1}(x_i)-F_{e2}(y_i)\mid \tag{3.29}$$

여기서 $F_{e1}(x_i)$와 $F_{e2}(y_i)$는 자료수가 각각 N_1과 N_2인 두 표본자료의 경험적 누가분포함수이며, x_i와 y_i는 각각의 표본자료를 작은 값부터 크기순으로 재정열한 자료값이다.

식 (3.29)의 검정통계량을 이용하여 유의수준 α에서 $D^*<d_{1-\alpha/2}(N^*)$인 경우 두 표본자료는 동일한 확률분포형을 갖는다고 할 수 있다. 여기서 $N^*=N_1N_2/(N_1+N_2)$로 정의되며 $d_{1-\alpha/2}(N^*)$은 새롭게 정의된 자료수 N^*에 의해 결정되는 Kolmogorov-Smirnov 통계량이다.

【예제 3.9】 인도교지점의 연최대홍수량자료의 정규분포 여부에 대한 Kolmogorov-Smirnov 검정 결과는 다음과 같다.

인도교 홍수자료를 정규분포라 가정하였을 때 추정된 매개변수는 위치매개변수=13772.98, 규모매개변수=7205.897(예제 3.8참조)이다. 표 8에 크기순으로 재정열된 홍수자료와 각각의 홍수자료에 대한 도시공식 i/(N+1)을 이용한 경험적 누가분포, 추정된 매개변수를 이용하여 구한 적합된 누가분포, 이 두 값의 차이 D를 나타내었다. 표에서 보는 바와 같이 최대차이값은 0.1075임을 알 수 있다. 따라서 유의수준 $\alpha=0.05$에서 $d_{1-\alpha/2}(N)=0.210$이므로 $D<d_{1-0.05/2}(40)$ 인 조건을 만족하여 정규분포라는 귀무가설은 받아들여진다.

KOLMOGOROV-SMIRNOV GOODNESS OF FIT TEST
TABULATED TEST VALUE=.210
COMPUTED TEST VALUE=.1075
SIGNIFICANCE LEVEL=.050
HYPOTHESIS OF GOOD FIT CANNOT BE REJECTED

【예제 3.10】 인도교지점의 연최대홍수자료는 표 1에서 보는 바와 같이 자료 결측기간을 제외하면 1918-1940년과 1952-1991년 2개의 자료기간으로 분리되어 있다. 이 두 기간의 자료가 동일한 분

표 7. Quantile Points $d_\beta(N)$ for the Kolmogorov-Smirnov Test(Miller, 1956).

Sample size	Cumulative Probability(β)				Sample size	Cumulative Probability(β)			
N	0.80	0.90	0.95	0.99	N	0.80	0.90	0.95	0.99
1	.900	.950	.975	.995	21	.226	.259	.287	.344
2	6.84	.776	.842	.929	22	.221	.252	.281	.337
3	.565	.636	.708	.829	23	.216	.247	.275	.330
4	.492	.565	.624	.734	24	.212	.242	.269	.323
5	.447	.509	.563	.669	25	.208	.238	.264	.317
6	.410	.468	.519	.617	26	.204	.233	.259	.311
7	.381	.436	.483	.576	27	.200	.229	.254	.305
8	.358	.410	.454	.542	28	.197	.225	.250	.300
9	.339	.387	.430	.513	29	.193	.221	.246	.295
10	.323	.369	.409	.489	30	.190	.218	.242	.290
11	.308	.352	.391	.468	31	.187	.214	.238	.285
12	.296	.338	.375	.449	32	.184	.211	.234	.281
13	.285	.325	.361	.432	33	.182	.208	.231	.277
14	.275	.314	.349	.418	34	.179	.205	.227	.273
15	.266	.304	.338	.404	35	.177	.202	.224	.269
16	.258	.295	.327	.392	36	.174	.199	.221	.265
17	.250	.286	.318	.391	37	.172	.196	.218	.222
18	.244	.279	.309	.371	38	.170	.194	.215	.258
19	.237	.271	.301	.361	39	.168	.191	.213	.255
20	.232	.265	.294	.352	40	.165	.189	.210	.252
large N A_β	0.0256	0.05256	0.11282	0.28464	large N A_β	0.00256	0.05256	0.11282	0.28464

$$d_\beta(N) = \sqrt{\frac{\ln[1/(1-\beta)]}{2N}} \; \frac{0.16693}{N} A_\beta N^{-0.5}$$

표 8. Kolmogorov-Smirnov 검정방법의 계산예

No.	홍수량	경험적 누가분포	적합된 누가분포	D	No.	홍수량	경험적 누가분포	적합된 누가분포	D
1	2838.8	.0646	.0244	.0402	21	13200.2	.4686	.5122	.0439
2	3855.7	.0844	.0488	.0356	22	13282.3	.4729	.5366	.0637
3	4420.4	.0972	.0732	.0240	23	13750.2	.4987	.5610	.0622
4	5310.1	.1201	.0976	.0225	24	13987.2	.5119	.5854	.0735
5	5883.8	.1368	.1220	.0148	25	14193.4	.5233	.6098	.3865
6	5971.3	.1395	.1463	.0069	26	14255.4	.5267	.6341	.1075
7	6083.1	.1429	.1707	.0278	27	14811.6	.5573	.6585	.1012
8	6585.9	.1593	.1951	.0358	28	15531.4	.5964	.6829	.0865
9	6819.1	.1673	.2195	.0522	29	15922.2	.6172	.7073	.0901
10	6824.5	.1975	.2439	.0764	30	16365.3	.6405	.7317	.0912
11	9712.1	.2865	.2683	.0182	31	17517.6	.6984	.7561	.0577
12	9863.6	.2937	.2927	.0010	32	18874.6	.7605	.7805	.0200
13	9997.5	.3002	.3171	.0169	33	19048.9	.7680	.8049	.0369
14	10438.4	.3218	.3415	.0197	34	19399.8	.7826	.8293	.0467
15	11319.5	.3667	.3659	.0009	35	22338.0	.8827	.8537	.0290
16	11809.2	.3926	.3902	.0024	36	24430.0	.9304	.8780	.0524
17	12210.2	.4142	.4146	.0005	37	26431.3	.9605	.9024	.0581
18	12791.1	.4458	.4390	.0068	38	27409.4	.9708	.9268	.0440
19	12791.1	.4458	.4634	.0176	39	28836.3	.9817	.9512	.0305
20	12822.5	.4475	.4878	.0403	40	32986.1	.9962	.9756	.0206

포형을 갖는지 Kolmogorov-Smirnov 검정방법을 이용한 결과는 다음과 같다. 2개의 자료별로 도시공식 i/(N+1)을 사용하여 구한 경험적 누가분포 함수는 그림 1과 같다. 이 두가지 경험적 누가분포 곡선에서의 최대 차이값 $D^*=0.1829$임을 알 수 있다. 2개의 자료기간 $N_1=23$, $N_2=40$이므로 $N^*=14$이다. 유의수준 $\alpha=0.05$에서의 $d_{1-\alpha/2}/2(N^*)=0.349$이다. 따라서 $D^*<d_{1-\alpha/2}(N^*)$을 만족하므로 2개의 자료기간이 같은 분포형을 갖는다는 귀무가설은 유의수준 $\alpha=0.05$에서 받아들여진다고 할 수 있다.

KOLMOGOROV-SMIRNOV TEST FOR TWO SAMPLES
TABULATED TEST VALUE=.349
COMPUTED TEST VALUE=.1829
SIGNIFICANCE LEVEL=.050 HYPOTHESIS OF THE SAME DISTRIBUTION CANNOT BE REJECTED

3.3.3 Cramer von Mises Test

이 검정방법도 Kolmogorov-Smirnov 검정방법과 마찬가지로 누가분포함수에 대하여 검정하는 방법이다. 임의의 표본자료 $Y_1, \cdots Y_N$의 누가분포함수가 $F(y;\underline{\theta})$라고 할 때 Cramer von Mises 검정통계량은 다음과 같이 주어진다(Thompson, 1966).

$$W=\frac{1}{12N}+\sum_{i=1}^{N}\left[F(y_i;\underline{\theta})-\frac{2i-1}{2N}\right]^2 \qquad (3.30)$$

유의수 α에서 $W>w_{1-\alpha}(N)$를 만족하면 선정한 확률분포형이 표본자료에 대하여 적합하다고 할 수 있다. 여기서 $w_{1-\alpha}(N)$은 Cramer von Mises 통계값으로 α와 N의 함수이며, 자료수가 $N>20/\sqrt{\alpha}$이면 통계값은 α만의 함수로 표 9와 같다.

【예제 3.11】 인도교지점 연최대홍수량자료가 정규분포인가 Cramer von Mises 검정방법을 이용하여 알아보자.

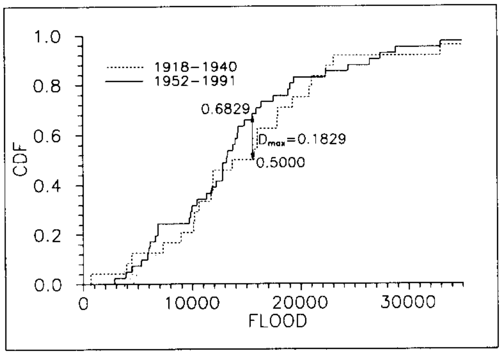

그림 1. Kolmogorov-Smirnov검정방법에서의 경험적 누가분포

표 9. Quantile Points w_β for the Cramer-von Mises Test(Stephens and Maag, 1969).

Sample size	Cumulative Probability(β)		
N	0.90	.424	.550
2	.343	.423	.622
3	.337	.435	.653
4	.341	.441	.671
5	.343	.451	.698
8	.346	.454	.707
10	.347	.459	.724
20	.349	.451	.743
∞	.347		

인도교 홍수자료를 정규분포라 가정하였을 때 추정된 매개변수는 위치매개변수=13772.98, 규모매개변수=7205.897(예제 3.8참조)이다. 표 10에는 순위홍수량과 이에 대응하는 누가분포, 그리고 식(3.30)을 이용한 Cramer von Mises검정통계량의 합을 나타내었다. 마지막 값이 W에 해당하며, 그 값은 W=0.11095이다. 유의수준 $\alpha=0.05$에서의 $w_{1-0.05}(40)=0.461$이므로 $W>w_{1-\alpha}(N)$인 조건을 만족하므로 정규분포라는 귀무가설은 성립한다.

CRAMER-VON MISES GOODNESS OF FIT TEST
TABULATED TEST VALUE=.461
COMPUTED TEST VALUE=.11095
SIGNIFICANCE LEVEL=.050
HYPOTHESIS OF GOOD FIT CANNOT BE REJECTED

표 10. Cramer von Mises검정방법 계산예

No.	순위홍수량	누가분포	ΣW	No.	순위홍수량	누가분포	ΣW
1	2838.8	.064584	.00480	21	13200.2	.468322	.02294
2	3855.7	.084369	.00699	22	13282.3	.472855	.02712
3	4420.4	.097160	.00819	23	13750.2	.498739	.03119
4	5310.1	.120110	.00926	24	13987.2	.511859	.03691
5	5883.8	.136797	.00985	25	14193.4	.523263	.04487
6	5971.3	.139475	.00985	26	14255.4	.526689	.05715
7	6083.1	.142949	.01023	27	14811.6	.557303	.06822
8	6585.9	.159288	.01103	28	15531.4	.596395	.07652
9	6819.1	.167266	.01308	29	15922.2	.617247	.08559
10	6824.5	.167454	.01798	30	16365.3	.640483	.09500
11	9712.1	.286530	.01856	31	17517.6	.698350	.09912
12	9863.6	.293728	.01860	32	18874.6	.760521	.09985
13	9997.5	.300159	.01875	33	19048.9	.767967	.10183
14	10438.4	.321769	.01900	34	19399.8	.782559	.10485
15	11319.5	.366747	.01902	35	22338.0	.882704	.10526
16	11809.2	.392610	.01904	36	24430.0	.930421	.10710
17	12210.2	.414153	.01905	37	26431.3	.960512	.10940
18	12791.1	.445808	.01911	38	27409.4	.970781	.11051
19	12791.1	.445808	.01939	39	28836.3	.981710	.11088
20	12822.5	.447531	.02099	40	32986.1	.996165	.11095

맺음말

이번 강좌에서는 수공학에서 많이 사용하는 검정방법에 대해서 인도교지점의 연최대홍수량자료를 이용하여 설명하였다. PPCC검정방법에 대해서는 정규분포인 경우에 대해서만 설명을 하였지만 이 방법은 정규분포외에도 2개의 매개변수를 갖는 대수정규분포, Gumbel분포에 대해서도 확장 적용되었으며(Vogel, 1986), 2개의 매개변수를 갖는 Weibull분포형에도 적용 가능한 것으로 저자는 확인하였다. 앞으로는 PPCC검정방법이 널리 쓰이기를 기대하며, 다음 강좌에서는 수문자료의 특성상 나타날 수 있는 자료의 경향(trend)과 변동(change)을 측정하는 방법에 대해 설명하기로 한다.

참 고 문 헌

Anderson, R.L.(1942). "Distribution of the serial correlation coefficient." Annals of Mathematical Statistics, Vol. 13, No. 1, pp. 1-13.

Benjamin, J.R. and Cornell, C.A.(1970). Probability, Statistics, and Decision for Civil Engineers, McGraw Hill, New York.

Clake, R.T.(1973). Mathematical Models in Hydrology, Irrigation and Drainage Paper 19, Food and Agriculture Organization, Rome.

Filliben, J.J.(1975). "The probability plot correlation coefficient test for normality." Technometrics, Vol. 17, No. 1, pp. 111-117.

Keeping, E.S.(1966). "Distribution free methods in statistics" in Proceedings of Hydrology Symposium No. 5, McGill University, Canada.

Miller, L.H.(1956). "Table of percentage points of Kolmogorov statistics." American Statistical Association, pp. 111-121.

Pearson, E.S. and Hartley, H.O.(1966). Biometrika Tables for Statisticians Vol. I, Cambridge.

Royston, J.P.(1982). "An extension of Shapiro and Wilk's W test for normality to large samples." Applied Statistics, Vol. 31, pp. 115–124.

Shapiro, S.S. and Wilk, M.B.(1965). "An analysis of variance test for normality." Biometrika, 52, pp. 591–611.

Snedecor, G.W. and Cochran, W.G.(1980). Statistical Methods. The Iowa State University Press, Ames, Iowa.

Stephens, M.A. and Maag, U.R.(1968). "Further percentage points for W." Biometrika, Vol. 55, No. 2, pp. 428–430.

Thompson, R.(1966). "Bias of the one-sample Cramer von Mises test." American Statistical Association, Vol. 61, pp. 246–247.

Yevjevich, V.(1972a). Probability and Statistics in Hydrology, Water Resources Publications, Fort Collins, Colorado.

Yevjevich, V.(1972b). Stochastic Processes in Hydrology, Water Resources Publications, Fort Collins, Colorado.

Vogel, R.M.(1986). "The probability plot correlation coefficient test for the normal, lognormal, and Gumbel distributional hypothesis", Water Resources Research, Vol. 22, No. 4, pp. 587–590. ♧