응용통계연구 제 12권 2호, 1999년, pp. 553-561

상관계수에 대한 모수적 추론 : 대안적 방법*

허명회[1] 김미경[2]

요 약

이변량 정규분포의 상관계수 ρ에 대한 검정 및 신뢰구간을 구하는 모수적 방법으로서 Fisher의 z 변환과 해당하는 점근적 분포가 널리 쓰이고 있다. 본 연구에서는 이에 대한 대안으로서 직교변환과 F 분포를 활용하는 방법을 제시한다. 후자의 방법이 전자와 비교하여 사실상 대등하면서도 설명은 오히려 쉬우므로 통계학 교육에 더 적합하다고 생각한다. 또한, 시험적으로, $H_0 : \rho = \rho_0$에 대한 모수적 임의화 검정법을 제안한다.

1. 서론

(X_1, X_2)에 대한 n개의 관측 $(x_{11}, x_{12}), \cdots, (x_{n1}, x_{n2})$가 모상관계수가 ρ인 이변량 정규분포 $\mathrm{BN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$로부터 생성되었다고 전제하고 표본상관계수를 r이라고 하자. 이때, ρ에 대한 가설

$$H_0 \ : \ \rho = \rho_0 \quad 대 \quad H_1 \ : \ \rho > \rho_0, \qquad \rho_0는\ -1과\ 1사이의\ 既知\ 定數 \tag{1}$$

의 검정방법과 ρ에 대한 신뢰구간을 구하는 방법에 대하여 생각하기로 한다.

가장 흔히 쓰이는 모수적 방법은 피셔 (R.A. Fisher)의 z 변환과 이에 부수된 대표본성질을 활용하는 것이다. 즉,

$$h(r) = 0.5 \log_e\{(1+r)/(1-r)\} \tag{2}$$

로 변환하고 이것의 점근적 표집성질인

$$\sqrt{n-3}\,\{h(r) - h(\rho)\} \ \sim \ N(0, 1) \tag{3}$$

을 이용하는 방법이다 (Fisher, 1958, Chapter VI; Hotelling, 1953). 그러나, 초급 통계학 강의에서는, 변환식 (2)와 분포적 성질 (3)을 유도하고 설명하기가 쉽지 않다는 데 누구나 동의할 것이다. 다음 절에서 이에 대한 대안으로서 한 모수적 방법을 제시한다.

2. 새로운 모수적 방법

$Y_1 = X_1/\sigma_1$, $Y_2 = X_2/\sigma_2$로 변환하면 (Y_1, Y_2)는 단위분산을 갖는 이변량 정규분포 $\mathrm{BN}(\nu_1, \nu_2, 1, 1, \rho)$를 따른다 (여기서 $\nu_1 = \mu_1/\sigma_1$, $\nu_2 = \mu_2/\sigma_2$). 그리고 Y_1과 Y_2의 합 $U_1 = Y_1 + Y_2$와 차이 $U_2 = Y_1 - Y_2$는

* 본 연구는 한국학술진흥재단의 1998년도 자유공모과제 지원을 받아 수행되었습니다.

1) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 정경대학 통계학과, 교수
E-mail: stat420@kuccnx.korea.ac.kr

2) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계학과, 박사과정

$$\mathrm{BN}(\nu_1+\nu_2,\ \nu_1-\nu_2,\ 2(1+\rho),\ 2(1-\rho),\ 0)$$

을 따른다. 즉 U_1과 U_2는 무상관(no correlation)이므로, U_1와 U_2의 n개 임의관측 $(u_{11}, u_{12}), \cdots, (u_{n1}, u_{n2})$의 평균수정 제곱합들인

$$\sum_{i=1}^{n}(u_{i1}-\bar{u}_1)^2 \equiv (n-1)\ s_{u_1}^2, \quad \sum_{i=1}^{n}(u_{i2}-\bar{u}_2)^2 \equiv (n-1)\ s_{u_2}^2$$

은 독립적으로 $2(1+\rho)\chi^2_{n-1}$분포와 $2(1-\rho)\chi^2_{n-1}$분포를 따르게 된다. 따라서

$$F \equiv \frac{s_{u_1}^2/(1+\rho)}{s_{u_2}^2/(1-\rho)} \ \sim\ F_{n-1,n-1} \tag{4}$$

이 유도된다. 이로부터, σ_1과 σ_2를 알고 있다면, (1)의 가설검정은

$$F \equiv \frac{s_{u_1}^2/(1+\rho_0)}{s_{u_2}^2/(1-\rho_0)} \ \sim\ F_{n-1,n-1} \tag{5}$$

에 의하여, 그리고 ρ에 대한 수준 $1-\alpha$의 신뢰구간은

$$F_{n-1,n-1}^{\alpha/2}\frac{1+\rho}{1-\rho} \ \le\ \frac{s_{u_1}^2}{s_{u_2}^2} \ \le\ F_{n-1,n-1}^{1-\alpha/2}\frac{1+\rho}{1-\rho} \tag{6}$$

로부터 얻을 수 있다.

그런데 (5)와 (6)이 미지의 모수 σ_1과 σ_2에 의존하므로 실제에서는 쓸모가 없다. σ_1과 σ_2 대신 이것들을 변수 X_1과 X_2의 표본표준편차 s_1과 s_2로 대치하면 식 (4)의 F는

$$\tilde{F} \equiv \frac{(1+r)/(1+\rho)}{(1-r)/(1-\rho)} \tag{7}$$

이 되는데, 더 이상 이것은 $F_{n-1,n-1}$분포를 정확히 따른다고는 말할 수 없다. 그렇지만 s_1과 s_2가 σ_1과 σ_2에 대한 일치 추정량이므로, 근사적으로는 $\tilde{F}$의 분포가 $F_{n-1,n-1}$, 또는 이것보다 다소 이완된 $F_{n-2,n-2}$이나 $F_{n-3,n-3}$에 가까울 것으로 유추할 수 있겠다.

다음 절에서 Monte Carlo 계산을 통해 그 정확성을 검토해 보기로 한다.

3. 몬테칼로 방법에 의한 평가

식 (7)에 log를 취하고 0.5배 해주면

$$z \equiv 0.5\ \log_e \tilde{F} = h(r) - h(\rho)$$

가 되므로 다음과 같이 Fisher의 방법과 본 연구에서 제시하는 대안적 방법을 대비시킬 수 있다.

	Fisher의 방법	대안적 방법 1, 2, 3
z의 근사 분포 :	$(n-3)^{-\frac{1}{2}}\,\mathrm{N}(0,1)$	$0.5\log_e F_{n-k,n-k},\ k=1,2,3$

$n = 10, 20, 40, 100$의 경우, z의 근사적 5%, 95%, 2.5%, 97.5% 분위수로서 두 방법이 제시하는 수치를 표 3.1과 표 3.2에 열거해 보았다. Fisher의 분위수와 특히 대안적 방법 2의 분위수가 거의 비슷하고 표본의 크기가 클수록 더욱 일치하는 경향을 볼 수 있다.

그렇다면 이들 근사적 분위수는 얼마만큼 실제 분위수에 가까운 것일까? 그것을 보기 위해, 상관계수 ρ (=0.00, 0.10, 0.25, 0.50, 0.75, 0.90)의 이변량 정규변량을 n (=10, 20, 40, 100)개 모의생성시키고 이런 과정을 999번씩 100회 반복하여 z의 몬테칼로 분위수를 추정해보았다. 그리고, 그 결과를 표 3.1과 표 3.2 안에 제시하였다.

표본크기 n이 크지 않은 경우에는, z의 경험적 5%, 95% 분위수와 2.5%, 97.5% 분위수가 모두 ρ가 커짐에 따라 대체로 증가함을 표 3.1과 표 3.2로부터 볼 수 있다. 즉, z가 정확한 추축량(pivot quantity)은 아닌 것이다. 단, 그 변화 폭이 크지 않고 특히 분위수간 거리(=95%분위수 - 5%분위수, 97.5%분위수 - 2.5%분위수)에 있어 Fisher의 방법과 대안적 방법 2의 분위수간 거리가 경험적 분위수간 거리보다 대체로 비슷하거나 크므로, 이들 근사적 방법들이 신뢰구간 및 양측 가설검정에서 다소 '보수적'인 특성을 가질 것으로 예상할 수 있다 (여기서 보수성은 신뢰구간의 경우 실제 포함확률이 명목적 수준을 초과함, 가설검정의 경우 제 1종 오류 확률이 규정된 수준에 미달함을 의미한다).

그러나 단측 가설검정에서는 두 모수적 근사 방법 모두에, 특히 표본크기가 작은 경우에, 우려할만한 문제가 있을 수 있다. 표 3.1과 표 3.2에서 근사적 방법이 제시하는 95% 분위수 및 97.5% 분위수가 큰 ρ에 있어서는 Monte Carlo 분위수에 상당히 못미치므로

$$H_0 : \rho = \rho_0 \quad \text{대} \quad H_1 : \rho > \rho_0, \qquad \rho_0\text{는 양의 旣知數} \tag{8}$$

의 가설검정에서는 근사적 방법이 실제보다 작은 p 값을 보고할 가능성이 있기 때문이다. 다음 절에서 수치예를 다루면서 이제까지 논의된 절차를 적용해보기로 하겠다.

4. 수치 예

Efron (1982)의 법무대학원 입학자료를 예로 들기로 한다. 이 자료는 15개 학교 입학생들의 평균 LSAT 점수와 평균 GPA로 구성되어 있으며 ($n = 15$), 두 변수간 상관계수 r은 0.776이다. 이로부터 Fisher의 방법과 대안적 방법 2를 따라 95% 신뢰구간을 구하여보자.

1) Fisher의 방법 : $h(r) = 1.035$이므로 $h(\rho)$에 관한 95% 신뢰구간은

$$1.035 - 1.96/\sqrt{12} \;\le\; h(\rho) \;\le\; 1.035 + 1.96/\sqrt{12}$$

즉 $0.469 \le h(\rho) \le 1.601$로부터 $0.437 \le \rho \le 0.922$가 된다.

2) 대안적 방법 2 : $F_{13,13}$분포의 2.5% 분위수, 97.5% 분위수가 각각 3.12^{-1} 및 3.12이고 $(1+r)/(1-r) = 7.929$이므로 ρ에 관한 95% 신뢰구간은

표 3.1: z의 5%, 95% 분위수 및 그 차이

5% 분위수 $z_{.05}$		$n=10$	$n=20$	$n=40$	$n=100$
Fisher		-0.622	-0.399	-0.270	-0.167
대안 1		-0.578	-0.387	-0.267	-0.166
2		-0.618	-0.398	-0.270	-0.167
3		-0.666	-0.410	-0.274	-0.168
Monte Carlo	$\rho=0.00$	-0.621 (0.003)	-0.400 (0.002)	-0.272 (0.001)	-0.169 (0.001)
	0.10	-0.613 (0.003)	-0.398 (0.002)	-0.268 (0.001)	-0.165 (0.001)
	0.25	-0.595 (0.003)	-0.392 (0.002)	-0.267 (0.001)	-0.165 (0.001)
	0.50	-0.583 (0.003)	-0.384 (0.002)	-0.263 (0.001)	-0.164 (0.001)
	0.75	-0.559 (0.003)	-0.376 (0.002)	-0.259 (0.001)	-0.162 (0.001)
	0.90	-0.541 (0.002)	-0.370 (0.002)	-0.257 (0.001)	-0.163 (0.001)
95% 분위수 $z_{.95}$		$n=10$	$n=20$	$n=40$	$n=100$
Fisher		0.622	0.399	0.270	0.167
대안 1		0.578	0.387	0.267	0.166
2		0.618	0.398	0.270	0.167
3		0.666	0.410	0.274	0.168
Monte Carlo	$\rho=0.00$	0.625 (0.003)	0.400 (0.002)	0.272 (0.001)	0.166 (0.001)
	0.10	0.622 (0.003)	0.404 (0.002)	0.273 (0.001)	0.169 (0.001)
	0.25	0.639 (0.003)	0.406 (0.002)	0.272 (0.001)	0.169 (0.001)
	0.50	0.642 (0.002)	0.411 (0.002)	0.275 (0.001)	0.170 (0.001)
	0.75	0.660 (0.002)	0.420 (0.002)	0.280 (0.001)	0.171 (0.001)
	0.90	0.669 (0.002)	0.424 (0.002)	0.280 (0.001)	0.173 (0.001)
$z_{.95} - z_{.05}$		$n=10$	$n=20$	$n=40$	$n=100$
Fisher		1.244	0.798	0.540	0.334
대안 1		1.156	0.774	0.534	0.332
2		1.236	0.796	0.540	0.334
3		1.332	0.820	0.548	0.334
Monte Carlo	$\rho=0.00$	1.246 (0.004)	0.800 (0.002)	0.544 (0.002)	0.335 (0.001)
	0.10	1.235 (0.004)	0.801 (0.002)	0.541 (0.001)	0.334 (0.001)
	0.25	1.234 (0.004)	0.798 (0.002)	0.539 (0.002)	0.333 (0.001)
	0.50	1.225 (0.003)	0.795 (0.002)	0.538 (0.001)	0.333 (0.001)
	0.75	1.219 (0.004)	0.796 (0.002)	0.539 (0.002)	0.333 (0.001)
	0.90	1.210 (0.003)	0.794 (0.002)	0.537 (0.002)	0.336 (0.001)

*( ) 안은 표준오차

표 3.2: z의 2.5%, 97.5% 분위수 및 그 차이

2.5% 분위수 $z_{.025}$		$n=10$	$n=20$	$n=40$	$n=100$
Fisher		-0.741	-0.475	-0.322	-0.199
대안 1		-0.696	-0.463	-0.319	-0.198
	2	-0.745	-0.477	-0.323	-0.199
	3	-0.804	-0.492	-0.327	-0.200
Monte Carlo	$\rho=0.00$	-0.752 (0.004)	-0.478 (0.002)	-0.327 (0.001)	-0.202 (0.001)
	0.10	-0.746 (0.004)	-0.478 (0.002)	-0.321 (0.001)	-0.200 (0.001)
	0.25	-0.727 (0.003)	-0.473 (0.002)	-0.319 (0.001)	-0.197 (0.001)
	0.50	-0.707 (0.003)	-0.463 (0.002)	-0.318 (0.001)	-0.196 (0.001)
	0.75	-0.684 (0.003)	-0.453 (0.002)	-0.313 (0.001)	-0.195 (0.001)
	0.90	-0.663 (0.003)	-0.447 (0.002)	-0.311 (0.001)	-0.196 (0.001)
97.5% 분위수 $z_{.975}$		$n=10$	$n=20$	$n=40$	$n=100$
Fisher		0.741	0.475	0.322	0.199
대안 1		0.696	0.463	0.319	0.198
	2	0.745	0.477	0.323	0.199
	3	0.804	0.492	0.327	0.200
Monte Carlo	$\rho=0.00$	0.754 (0.004)	0.479 (0.002)	0.325 (0.001)	0.199 (0.001)
	0.10	0.751 (0.004)	0.484 (0.002)	0.327 (0.001)	0.201 (0.001)
	0.25	0.765 (0.003)	0.484 (0.002)	0.324 (0.001)	0.201 (0.001)
	0.50	0.770 (0.003)	0.491 (0.002)	0.329 (0.001)	0.201 (0.001)
	0.75	0.790 (0.003)	0.499 (0.002)	0.335 (0.001)	0.203 (0.001)
	0.90	0.795 (0.003)	0.502 (0.002)	0.334 (0.001)	0.205 (0.001)
$z_{.975} - z_{.025}$		$n=10$	$n=20$	$n=40$	$n=100$
Fisher		1.482	0.950	0.644	0.398
대안 1		1.392	0.926	0.638	0.396
	2	1.490	0.954	0.646	0.398
	3	1.608	0.984	0.654	0.400
Monte Carlo	$\rho=0.00$	1.506 (0.005)	0.957 (0.003)	0.652 (0.002)	0.401 (0.001)
	0.10	1.497 (0.005)	0.962 (0.003)	0.648 (0.002)	0.401 (0.001)
	0.25	1.492 (0.005)	0.957 (0.003)	0.644 (0.002)	0.398 (0.001)
	0.50	1.478 (0.004)	0.954 (0.003)	0.647 (0.002)	0.397 (0.001)
	0.75	1.474 (0.005)	0.952 (0.003)	0.647 (0.002)	0.398 (0.001)
	0.90	1.458 (0.004)	0.950 (0.003)	0.644 (0.002)	0.401 (0.001)

*( ) 안은 표준오차

$$3.12^{-1}\frac{1+\rho}{1-\rho} \le 7.929 \le 3.12\frac{1+\rho}{1-\rho},$$
즉 $0.435 \le \rho \le 0.923$이다

이제는 $\rho_0 = 0.5$로 주어졌다고 하고 (8)의 단측검정 문제를 생각하여 보자.

$$z = h(r) - h(\rho_0) = 0.486$$

이므로 Fisher의 방법과 대안적 방법 2에 따른 p 값은 다음과 같이 산출된다.

1) Fisher의 방법 : $\sqrt{12}\,z = 1.684$이므로
p 값 $= 1 - \Phi(1.684) = 0.0462$.

2) 대안적 방법 2 : $\tilde{F} \equiv \dfrac{(1+r)/(1+\rho_0)}{(1-r)/(1-\rho_0)} = 2.643$이므로
p 값 $= 1 - F_{13,13}(2.643) = 0.0458$.

그러나 몬테칼로 방법 (반복수 99,900)으로 p 값을 추정해보면

$$\text{p 값} = 0.0522 \ (s.e. = 0.0007)$$

이다. 그러므로 Fisher의 방법과 대안적 방법 2가 모두 유의수준 5%에서 영가설을 기각하지만, Monte Carlo 방법으로 평가해본 결과 실제적인 p 값이 5%를 초과하기 때문에 영가설을 기각해서는 안되는 경우인 것이다.

5. 임의화 검정의 가능성

상관계수에 관한 가설검정에서 $\rho_0 = 0$인 경우 임의화 검정(randomization test)은 잘 알려져 있다 (Good, 1994, Chapter 7). 그러나 $\rho_0 \neq 0$인 경우의 임의화 검정은 아직 개발되어 있지 않은 듯하다. 이제 모수적 형태의 임의화 검정법을 제안하기로 한다.

(X_1, X_2)에 대한 n개의 관측 $\{(x_{11}, x_{12}), \cdots, (x_{n1}, x_{n2})\}$가 모상관계수가 ρ인 이변량 정규분포

$$\mathrm{BN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$$

로부터 생성되었다고 전제하자. 그러면 이들의 결합확률밀도가

$$\prod_{i=1}^{n} f_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho}(x_{i1}, x_{i2}) \propto (\sigma_1\sigma_2\sqrt{1-\rho^2})^{-n/2}\exp\left[-(1-\rho^2)^{-1}Q/2\right] \tag{9}$$

로 주어진다. 여기서

$$Q = \sum_{i=1}^{n}(x_{i1}-\mu_1)^2/\sigma_1^2 - 2\rho\sum_{i=1}^{n}(x_{i1}-\mu_1)/\sigma_1\cdot(x_{i2}-\mu_2)/\sigma_2 + \sum_{i=1}^{n}(x_{i2}-\mu_2)^2/\sigma_2^2$$

이다. 결합확률밀도 (9)는, $\mu_1 = \bar{x}_1,\ \mu_2 = \bar{x}_2,\ \sigma_1^2 = s_1^2,\ \sigma_2^2 = s_2^2$일 때 즉 장애모수들을 그것에 대한 일치추정치로 대체해 놓았을 때, 일종의 지수적 가중치인

$$w_\rho(r\,;n) \equiv \exp\left[\,(1-\rho^2)^{-1}\rho\cdot(n-1)\,r\,\right]$$

에 비례하게 된다. 따라서 (1)의 가설검정문제

$$H_0 : \rho = \rho_0 \quad \text{대} \quad H_1 : \rho > \rho_0, \qquad \rho_0\text{는 } -1\text{과 } 1\text{사이의 旣知 定數}$$

에 대하여 다음과 같은 임의화 검정의 몬테칼로(Monte Carlo) 절차를 제안한다.

단계 1: 원래의 관측 표본 $S = \{(x_{11}, x_{12}), \cdots, (x_{n1}, x_{n2})\}$로 부터 상관계수 r_0를 산출한다.

단계 2: 변량 X_2의 n개 관측값을 임의순열화하여 새 표본

$$S^* = \{(x_{11}, x_{s_1 2}), \cdots, (x_{n1}, x_{s_n 2})\}$$

를 생성시킨다 (여기서 $(s_1, \cdots, s_n)$은 $(1, \cdots, n)$의 임의순열이다). 그리고 상관계수 r_s^*와 해당하는 $w_{\rho_0}(r_s^*; n)$을 계산한다.

단계 3: 단계 2를 N번 독립적으로 반복하여 다음과 같이 p 값을 추정한다 (모든 순열을 취하지 않는 이유는 계산이 워낙 방대하기 때문이다).

$$P_{\text{random}} = \frac{\sum_s w_{\rho_0}(r_s^*; n)\cdot 1(r_s^* \geq r_0)}{\sum_s w_{\rho_0}(r_s^*; n)},$$

여기서 $1(u \geq 0)$은 $u \geq 0$인 경우 1, 그렇지 않은 경우 0의 값을 갖도록 정의되는 지표함수이다.

임의화검정으로부터의 p 값인 P_{random}의 표준오차를 구하기 위하여는 앞의 알고리즘을 M번 반복하면 된다. 다음은 Efron의 법무대학원 자료에 $N = 10000,\ M = 20$의 몬테칼로 절차를 거쳐 얻은 결과이다 ($\rho_0 = 0.5$).

$$P_{\text{random}} = 0.0862, \quad s.e. = 0.0046$$

6. 맺음말

앞 절에서 제안된 임의화 검정법은 모수적 형태의 가중치를 사용하므로 임의화 검정법으로는 특이하게 모수적이라고 하겠다. 그러나 피셔의 근사적 방법이나 2절에서 제안된 새로운 모수적 방법과 다른 점은 이 임의화 검정법이 표본자료에 조건화되어 있다는 점이다.

특히 상관계수 ρ에 대한 추론과 관련하여서 장애모수라고 할 수 있는 μ_1, μ_2와 σ_1^2, σ_2^2을 표본평균과 표본분산으로 대체시키므로 주어진 자료로부터의 직접적인 추론이 가능하다. 그러나 아직은 시험적 제안 단계이므로 앞으로 이에 대한 이론적 연구가 요구된다.

참고문헌

[1] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics.

[2] Fisher, R. A. (1958). *Statistical Methods for Research Workers*, Thirteenth Edition. Hafner Publishing.

[3] Good, P. (1994). *Permutation Test*, Springer-Verlag.

[4] Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of Royal Statistical Society*, Series B, 15, 193-232.

[1999년 1월 접수, 1999년 4월 최종수정]

Alternative Method of Parametric Inference for Correlation Coefficient*

Myung-Hoe Huh[1)] Mi-Kyung Kim[2)]

ABSTRACT

For the correlation coefficient of bivariate normal distribution, Fisher's z transform is frequently used with related asymptotic distribution for inferential purpose. However, it is not easy to explain the underlying theory or idea of Fisher's method in undergraduate classes.

In this study, we propose another inferential method based on orthogonal transform and the F-distribution, which can be described as follows. Let

$$\tilde{F} \equiv \frac{(1+r)/(1+\rho)}{(1-r)/(1-\rho)},$$

where r is the sample product-moment correlation coefficient of the random sample of size n from the bivariate normal distribution with population correlation coefficient ρ. We argue that $\tilde{F}$ follows $F_{n-2,n-2}$-distribution approximately. Then we derive a testing method for $H_0 : \rho = \rho_0$ where ρ_0 may not be zero and develop a confidence interval for ρ. Our new method is more readable, while its performance is about the same as Fisher's method. Also, we propose a randomization testing method for $H_0 : \rho = \rho_0$ where $\rho_0 \neq 0$.

* The authors wish to acknowledge the financial support of the Korea Research Foundation made in the program year of 1998.

1) Professor. Dept. of Statistics, Korea University. Anam-dong 5-1, Seoul 136-701, Korea.
E-mail: stat420@kuccnx.korea.ac.kr

2) Graduate Student. Dept. of Statistics, Korea University.