The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.4.645
27(4), 645–653

Reproducibility of Hypothesis Testing and Confidence Interval

Myung-Hoe Huh[a,1]

[a]Department of Statistics, Korea University

(Received June 9, 2014; Revised July 16, 2014; Accepted July 17, 2014)

Abstract

P-value is the probability of observing a current sample and possibly other samples departing equally or more extremely from the null hypothesis toward postulated alternative hypothesis. When *p*-value is less than a certain level called $\alpha(= 0.05)$, researchers claim that the alternative hypothesis is supported empirically. Unfortunately, some findings discovered in that way are not reproducible, partly because the *p*-value itself is a statistic vulnerable to random variation. Boos and Stefanski (2011) suggests calculating the upper limit of *p*-value in hypothesis testing, using a bootstrap predictive distribution. To determine the sample size of a replication study, this study proposes thought experiments by simulating boosted bootstrap samples of different sizes from given observations. The method is illustrated for the cases of two-group comparison and multiple linear regression. This study also addresses the reproducibility of the points in the given 95% confidence interval. Numerical examples show that the center point is covered by 95% confidence intervals generated from bootstrap resamples. However, end points are covered with a 50% chance. Hence this study draws the graph of the reproducibility rate for each parameter in the confidence interval.

Keywords: Reproducibility, hypothesis testing, *p*-value, bootstrap method, confidence interval.

1. 연구 배경과 목적

통계적 검정은 실증적 과학에서 광범위하게 활용되고 있고 대부분의 경우 *p*-값을 유의수준 α와 비교하여 연구가설이 지지되는가를 확인한다. 그러나 유의성이 확인된 가설이 재현성 연구에서는 부정되는 경우가 드물지 않다.

그 이유 중 하나는 *p*-값 자체가 임의적 수치이므로 확률변동 하에 있는데 이것의 변동성이 상당히 크기 때문이다. 수리적으로 알려진바, 영가설 하에서 연속적 검정 통계량이 사용된 경우 *p*-값은 균일분포를 따른다. 영가설이 유효하지 않은 상황에서 *p*-값의 행태는 붓스트랩 방법으로 파악될 수 있는데, Boos와 Stefanski (2011)의 방법은 H_0 대 H_1의 검정에서 통계량 T에 의한 *p*-값의 예측분포를 다음과 같이 구하는 것이다.

1) 관측자료 $x_1, \ldots, x_n$으로부터 통계량 T와 이 통계량의 *p*-값 p_{obs}를 산출한다.

[1]Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr

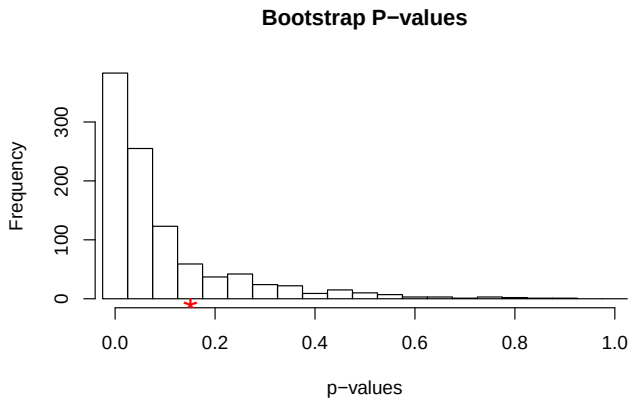

Figure 1.1. Bootstrap predictive distribution of the p-value for comparing two means

2) 붓스트랩 재표본 $x_1^*, \ldots, x_n^*$로부터 통계량 T^*와 이 통계량의 p-값 p^*를 산출한다. 이런 일을 독립적으로 반복하여 N개의 p^* 값을 생성해내고 이들을 $p_1^*, \ldots, p_N^*$로 표기한다.

3) $p_1^*, \ldots, p_N^*$의 경험적 분포를 히스토그램으로 파악하고 부가적으로 상위 20% 분위수 $p_{0.2}^*$를 산출한다 (또는 상위 10% 분위수 $p_{0.1}^*$).

앞의 단계 3에서 상위 20% 또는 상위 10% 분위수를 보는 이유는 그것이 검정력 80% 또는 검정력 90%와 관련이 되기 때문인데 이에 대하여는 곧 논의될 것이다.

간단한 예로서, 2개 평균을 비교하는 연구에서 표본 A와 표본 B가 관측되었다고 하자.

표본 A: 0.7, −1.6, −0.2, −1.2, −0.1, 3.4, 3.7, 0.8, 0.0, 2.0, 1.4 ($n_A = 11$)

표본 B: 1.0, 0.8, 1.1, 0.1, −0.1, 4.4, 5.5, 1.6, 4.6, 3.4 ($n_B = 10$)

표본 A와 표본 B가 각각 중심이 μ_A와 μ_B인 모집단에서 임의생성되고 두 모집단의 산포가 같다고 가정하자. 이 연구에서 설정된 가설은 $H_0 : \mu_A = \mu_B$ 대 $H_1 : \mu_A < \mu_B$이다. 2개 표본의 관측 결과는 $\bar{x}_A = 0.81$, $\bar{x}_B = 2.24$이고 차이는 $\bar{d} = -1.43$이다. 그리고 t-검정으로 계산된 p-값은 $0.048 (= p_{obs})$이다. 따라서 유의수준 $\alpha = 0.05$에서 대안가설 H_1이 지지된다.

이 연구에 대한 재현 실험에서 가설 H_1이 재확인될 수 있을까? p-값에 대한 붓스트랩 예측분포를 만들어 이에 대한 답을 할 수 있다. 이 경우는 관측자료가 독립표본 2개의 합이므로 붓스트랩 재표본도 A로부터의 재표본과 B로부터의 재표본으로 형성된다.

Figure 1.1은 p^*의 붓스트랩 반복 값 1,000개의 히스토그램인데 $p^* < 0.05$의 비율은 54.8 (±3.1)%에 불과하다. 즉 재현율(reproducibility probability; Shao와 Chow (2002), Goodman (1992))은 60%를 넘지 못한다.

이와 같이 재현성의 관점에서는 가설검증을 p-값 p_{obs}으로만 하는 것은 위험하다. p_{obs}를 보완하는 정보로 p^* 분포의 상위 20% 분위수 $p_{0.2}^*$를 활용할 필요가 있다. 앞의 사례에 돌아가 보자.

Figure 1.1에서 $p_{0.2}^*$는 기호 ' $*$ '로 표시되어 있는데 그 값은 0.151이다. 즉 관측 p-값 p_{obs}는 0.048이지만 이 값은 쉽게 0.151까지 커질 수 있음을 의미한다. 따라서 이 연구에서 지지된 가설 H_1의 재현성은

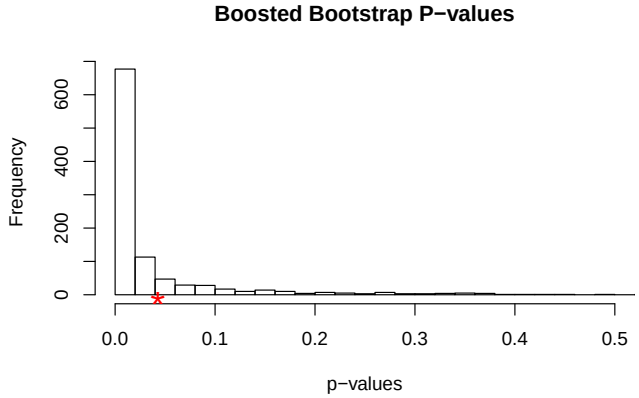

Figure 2.1. Bootstrap predictive distribution of the p-value from double size studies

의심스러울 수밖에 없다. 상위 20% 분위수 $p^*_{0.2}$ 대신 상위 10% 분위수 $p^*_{0.1}$를 쓸 수도 있을 것이다. 이 사례에서 $p^*_{0.1}$는 0.279이다.

재현성 연구에 나타날 p-값의 행태를 예측하는 이런 방법은 연구 규모를 임의로 할 수 있다는 데 장점이 있다. 다음 절에서 이에 대하여 다룬다. 제안 방법을 적용하면, 재현성 연구의 규모가 기존 연구의 k배인 경우, 기존 연구에서 지지된 가설이 재확인되거나 반박될 가능성을 미리 알 수 있다.

2. p-값의 예측분포

재현성 연구에서 표본의 크기를 n'으로 하자. 이하 n'을 현재 크기인 n의 $k(=1,2,3,\ldots)$배로 두겠지만 k가 꼭 정수여야 할 필요는 없다. 다만 $n'(=kn)$이 정수임을 가정한다. 다음 방법으로 k배 규모의 재현성 연구에서 p-값의 분포를 예측할 수 있다.

1) 관측표본 $x_1,\ldots,x_n$을 복원 임의추출하여 붓스트랩 재표본 $x^*_1,\ldots,x^*_{n'}$을 확보하여 통계량 T^*와 이 통계량의 p-값 p^*를 산출한다. 이런 일을 반복하여 N개의 p^* 값을 생성해내고 이들을 $p^*_1,\ldots,p^*_N$로 표기한다.

2) $p^*_1,\ldots,p^*_N$의 경험적 분포를 히스토그램으로 파악하고 상위 20% 분위수 $p^*_{0.2}$를 산출한다.

이와 같이 산출된 $p^*_{0.2}$가 α보다 작으면 검정력(test power)이 80% 이상이라고 볼 수 있다. 만약 그렇지 않다면 제시된 가설을 확인 또는 반박하기 위해서는 더 큰 표본이 필요하다. 이런 경우에는 k를 증가시켜 다시 앞의 절차를 적용해볼 필요가 있다.

예로서 앞 절의 2개 평균을 비교하는 사례에 돌아가서 표본 크기를 현 연구의 2배로 놓자. 즉 새 연구에서 표본 A는 크기가 22이고 표본 B는 크기가 20이 된다. Figure 2.1은 확대 붓스트랩 표본추출(boosted bootstrap sampling)로 얻어진 p-값의 예측 분포이다.

Figure 2.1에서 p-값의 상위 20% 분위수는 0.043이고 $\alpha(=0.05)$보다 작은 p-값의 비율, 즉 재현 확률은 81.5 (±2.5)%이다. 따라서 규모가 2배인 재현성 연구의 경우 검정력은 80% 정도로 볼 수 있다.

제안 방법은 분석 모형의 복잡도에 관계없이 활용이 가능하다. 다음 절은 다중선형회귀에 회귀 계수의 유의성 검정에 이 방법을 적용한 사례이다.

3. 다중선형회귀 사례

이 절에서의 분석 자료 "에어로빅 적합성"는 31명의 남자를 대상으로 측정된 7개 변수로 구성되어 있다. 반응변수는 oxy(oxygen uptake rate)이고 6개 설명변수는 age, wgt(weight), rtm(run time), rst(rest pulse), run(run pulse), max(maximum pulse)이다. 연구 가설이 'age, wgt, rtm, rst, run의 영향이 음이고 max의 영향은 양이다'라고 하자.

다음 표는 적합 회귀모형의 요약이다. 4개 변수 age, rtm, run, max가 유의하였고 ($\alpha = 0.05$), wgt와 rst는 유의하지 않았다.

	Coefficients	Std. Error	t-value	p-values
Intercept	102.93448	12.40326	8.299	
age	−0.22697	0.09984	−2.273	0.016
wgt	−0.07418	0.05459	−1.359	0.093
rtm	−2.62865	0.38456	−6.835	0.001
rst	−0.02153	0.06605	−0.326	0.374
run	−0.36963	0.11985	−3.084	0.003
max	0.30322	0.13650	2.221	0.018

이제 붓스트랩 방법으로 각 회귀계수의 p-값에 대한 예측분포를 살펴보기로 한다 (반복수 $N = 1000$). Figure 3.1이 작업 결과를 보여준다.

p-값의 상위 20% 분위수는 6개 설명변수에 대하여 각각 0.106 (age), 0.349 (wgt), 0.000 (rtm), 0.700 (rst), 0.045 (run), 0.156 (max)이다. 따라서 rtm과 run의 유의성은 확고한 편이지만 age와 max는 향후 번복될 가능성이 충분히 있다.

사고실험(thought experiment)의 규모를 현재 연구의 2배로 하자 ($n' = 62$). 그 결과로 얻는 6개 p-값의 상한(= 상위 20% 분위수)은 각각 0.062 (age), 0.244 (wgt), 0.000 (rtm), 0.583 (rst), 0.015 (run), 0.071 (max)이다. 따라서 age와 max의 상한이 α(= 0.05)를 초과하므로 2배 확대(boosting)로는 규모가 충분하지 않다.

이에 따라 재현성 연구를 현재 연구의 3배로 해볼 필요가 있고 ($n' = 93$), 그 결과로 얻는 p-값의 상위 20% 분위수는 6개 회귀계수에 대하여 각각 0.046 (age), 0.199 (wgt), 0.000 (rtm), 0.527 (rst), 0.010 (run), 0.054 (max)이다. Figure 3.2를 참조하라.

따라서 결론은 age 효과와 max 효과를 검증하기 위해 필요한 표본크기는 현재 규모의 최소 3배여야 한다는 것이다. wgt 효과와 rst 효과는 3배 규모로도 유의성이 검출되지 않을 것으로 보인다.

4. 신뢰구간 재현율

첫 번째 예로서, 2개 표본을 비교하는 1절의 사례에 돌아가 보자 ($n_A = 11$, $n_B = 10$). 이 자료에서 $\bar{x}_A = 0.81$, $\bar{x}_B = 2.24$이다. 그리고 $\mu_A - \mu_B (= \mu_D)$에 대한 95% t-신뢰구간은

$$-1.43 \pm 1.71 \quad \text{or} \quad (-3.14, 0.28)$$

이다. 이 구간 내 개별 파라미터 값들은 모두 동등한가? Bayesian 관점에서 이에 대한 답은 매우 명확히 '아니오(No)'이다. 재현성의 관점에서는 어떤가?

처리 A와 처리 B를 비교하는 "재현성" 연구가 미래에 같은 방법으로, 같은 크기로 반복적으로 시행된다

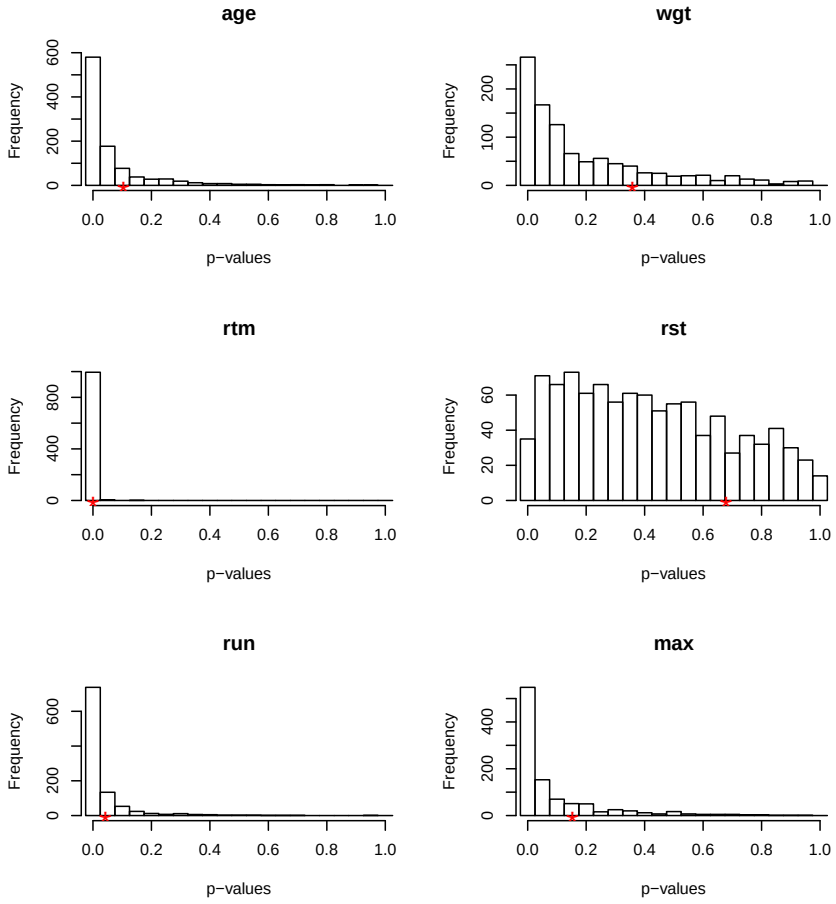

Figure 3.1. Bootstrap predictive distribution of the p-value of regression coefficients

고 하자. 그런 연구들 각각에서 μ_D에 대한 95% t-신뢰구간이 구해질 것이다. 구간 $(-3.14, 0.28)$의 개별 점 각각이 그런 구간들에 포함될 비율(= 재현율)을 구해 보자. 이 비율이 어떤 모수 값에서 50%라면 재현성이 반반 정도에 불과함을 의미한다.

많은 횟수의 재현성 연구는 현실적으로 있기 어렵다. 그러나 붓스트랩 표본추출로 이를 대신할 수 있다. 다음 알고리즘으로 재현율의 붓스트랩 추정값을 구할 것을 제안한다.

1) 크기 n의 관측표본 S로 관심 모수 θ에 대한 95% 신뢰구간을 구한다. 그 구간의 특정 점을 θ_0로 표기하자.
2) 관측표본 S로부터 $N(= 1000)$개의 독립적인 붓스트랩 재표본들 $S_1^*, \ldots, S_N^*$를 만들고 각 재표본(resample)에서 θ에 대한 95% 신뢰구간을 산출한다. 그리고 그런 구간들이 θ_0을 포함하는지의 여부를 조사하여 상대적 빈도를 산출한다. 이 값이 θ_0의 신뢰구간 재현율(추정치)이다.

Figure 4.1는 2개 표본을 비교한 앞의 사례에서 신뢰구간 재현율의 그래프이다. $1000(= N)$개씩의 붓

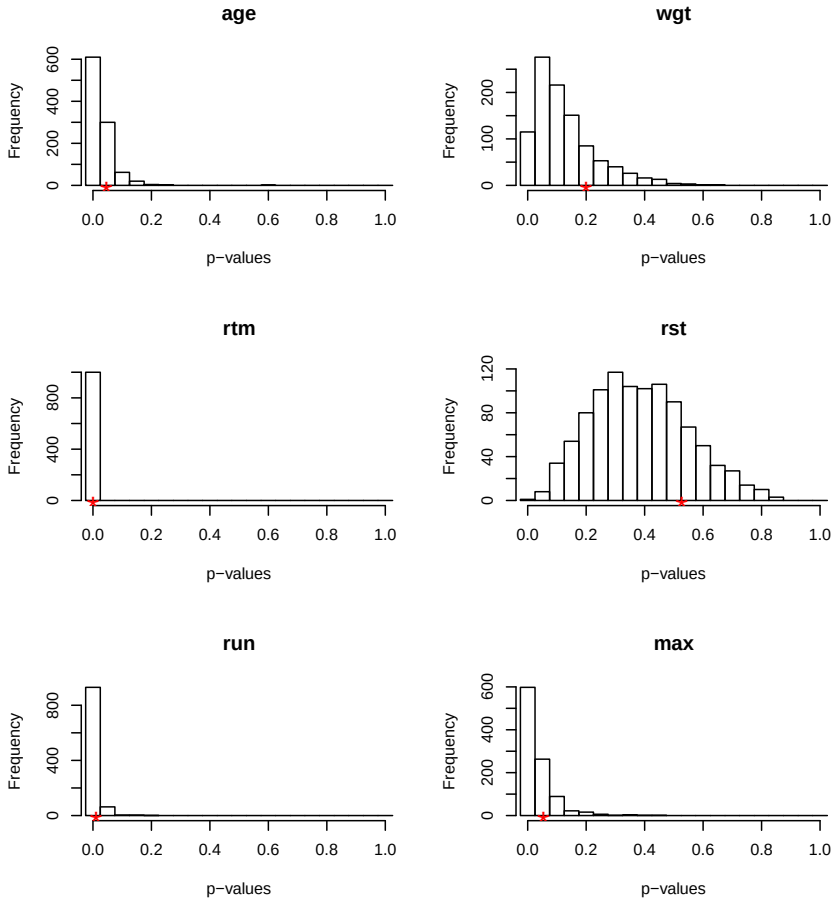

Figure 3.2. Bootstrap predictive distribution of the p-value of regression coefficients from triple size studies

스트랩 재표본이 사용되었다. 그림에 찍힌 점선은 관측표본 신뢰구간의 양끝 –3.14와 0.28에 위치해 있다. 재현율 함수는 확률밀도가 아니다. 각 점에서의 함수 값이 확률이다.

Figure 4.1의 그래프는 $\mu_D = -1.43$에서 정점(頂點)을 갖고 정점의 함수값은 대략 95%이다. 관측 신뢰구간 양끝의 재현율은 대략 50%이다. 이로써, 관측 신뢰구간 내 개별 점의 재현율은 50% 이상이다.

두 번째 예는 붓스트랩 신뢰구간의 재현율을 다룬다. 적용 자료는 고려대학교 2013년 '법과 통계학' 수강생 52명의 시험성적으로, 중간시험과 기말시험 간 상관계수는 $r = 0.489$이다. 붓스트랩 편향수정가속 방법(bootstrap bias-corrected and accelerated method; Efron, 1987)으로 구한 모상관계수 ρ에 대한 95% 신뢰구간은 $(0.259, 0.671)$이다. 이제 이 구간을 포함하는 충분한 너비의 구간 $(-0.25, 1)$의 개별 점에서 신뢰구간 재현율을 구해보기로 한다.

$N = 1000$개의 붓스트랩 재표본들을 생성시키고 각 재표본에서 ρ에 대한 붓스트랩 편향수정가속 신뢰구간을 구하여 개별 점이 이들 구간에 포함되는 비율, 즉 신뢰구간 재현율을 산출하였다. 이때 각 붓스트랩 재표본에서 ρ에 대한 신뢰구간을 구하는 과정에서 1000개의 붓스트랩 재표본이 사용되었다.

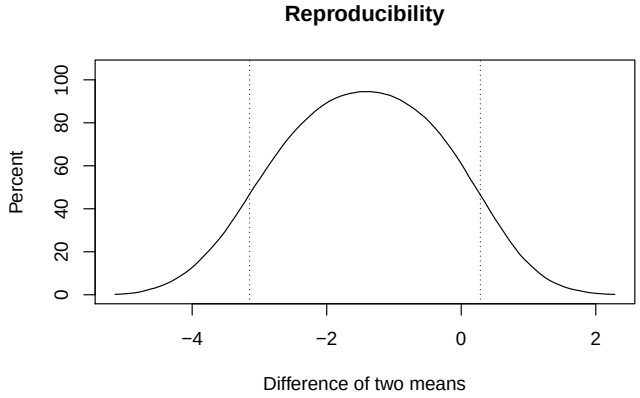

Figure 4.1. Replication rate of confidence interval for mean difference μ_D

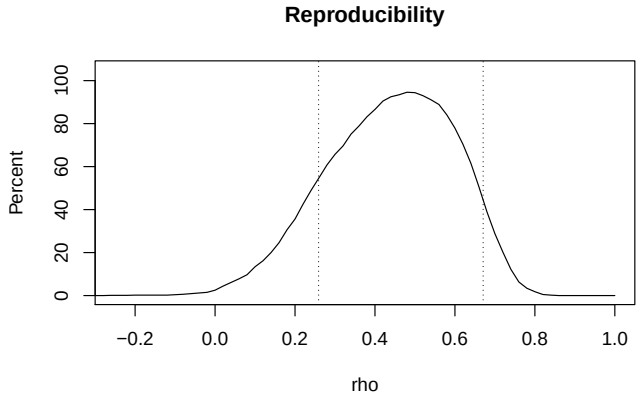

Figure 4.2. Replication rate of confidence interval for correlation coefficient ρ

Figure 4.2이 재현율의 그래프이다. 함수 형태가 비대칭적인 것을 볼 수 있는데 이는 상관계수의 표본추출분포가 비대칭적인 것과 관련이 있다. 관측 신뢰구간 $(0.259, 0.671)$의 양끝에서 재현율은 대략 50%이지만 좌단의 재현율이 우단의 재현율보다 다소 크다. 최고 재현율은 $\rho = 0.475$에서 대략 95%이다.

5. 맺음말

Hoenig과 Heisey (2001)의 지적대로 영가설이 기각되지 않은 경우 "관측 검정력(observed power)"은 50% 이하가 되므로 별 의미가 없다. 그러나 영가설이 기각되어 대안가설이 확인된 경우에는 "관측 검정력"이 나름대로 의미가 있다. 왜냐하면 이것이 충분히 크지 않은 경우 (예컨대 80% 미만인 경우) 향후 연구에서 재현성이 문제가 될 수 있기 때문이다.

Boos와 Stefanski (2011)의 연구는 모든 유의성 검정에서 p-값의 분포를 붓스트랩 방법으로 도출하고

상위 20% 분위수를 보고할 필요가 있음을 알렸다. 그 값이 $\alpha(= 0.05)$보다 작지 않은 경우에는 해당 통계량이 수준 α에서 유의하더라도 지지된 가설의 재현성은 확고하지 않다.

이 연구는 기존 연구에서 주장된 가설을 재확인 또는 반박하는 재현성 연구의 규모는 k배 확대 붓스트랩 표본추출로 찾을 것을 제안한다. 확대 인수 k를 정하기 위해서는 몇 차례 시행착오가 있을 수 있으나 계산적 부담은 크지 않다.

또한, 이 연구는 한 관측표본으로부터 얻어진 95% 신뢰구간 내 개별 점이 미래 연구의 신뢰구간에도 포함될 것인지 그 재현성을 붓스트랩 재표본들에서 평가하였다. 그 결과, 관측 신뢰구간의 중앙점은 재현율이 95%로 나오지만 신뢰구간 양끝 점은 재현율이 50%에 그침을 확인하였다.

References

Boos, D. D. and Stefanski, L. A. (2011). *P*-value precision and reproducibility, *The American Statistician*, **65**, 213–221.

Efron, B. (1987). Better bootstrap confidence intervals, *Journal of the American Statistical Association*, **82**, 171–185.

Goodman, S. N. (1992). A comment on replication, *p*-values and evidence, *Statistics in Medicine*, **11**. 875–879.

Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis, *The American Statistician*, **55**, 19–24.

Shao, J. and Chow, S.-C. (2002). Reproducibility probability in clinical trials, *Statistics in Medicine*, **21**, 1727–1742.

가설검정과 신뢰구간의 재현성

허명회[a,1]

[a]고려대학교 통계학과

(2014년 6월 9일 접수, 2014년 7월 16일 수정, 2014년 7월 17일 채택)

요 약

p-값은 관측 표본과 관측 결과보다 심하게 대안가설의 방향으로 영가설을 이탈하는 표본들이 영가설 하에서 갖는 확률이다. p-값이 일정 $\alpha(= 0.05)$보다 작게 나타나면 연구자는 대안가설이 지지된 것으로 본다. 그런 경우라고 하더라도 그의 가설이 향후 연구에서 번복될 수 있는데 그 이유는 p-값이 표본에 따라 변동하는 통계량이기 때문이다. Boos와 Stefanski (2011)는 붓스트랩 방법으로 p-값의 예측분포를 구할 수 있음을 보였다. 그들은 그 분포의 상위 10-20% 분위수가 α보다 작은가를 확인할 필요가 있음을 강조한다. 만약 그렇지 않은 경우에는 "지지"된 가설의 재현성이 문제될 수 있기 때문이다. 가설검정에서 일정 수준의 재현율을 확보하기 위해서는 표본의 증대가 요구된다. 이 연구는 k배 확대 붓스트랩 표본추출(boosted bootstrap sampling)로써 필요한 표본크기를 계산할 수 있음을 두 표본의 비교와 다중선형회귀의 수치 예에서 보인다. k 값을 정하기 위해서는 몇 차례 시행착오를 해야 하지만 계산적 부담은 크지 않다. 95% 신뢰구간은 독립적인 표본들로부터 같은 방식으로 산출되는 구간이 미지의 모수를 포함할 확률이 95%가 되도록 설정된다. 이 연구는 한 관측표본으로부터 얻어진 95% 신뢰구간 내 개별 점이 미래 연구의 신뢰구간에도 포함될 것인지 그 재현성을 붓스트랩 재표본들에서 평가한다. 이 연구는 개별 점에서 산출한 신뢰구간 재현율을 그래프로 보인다.

주요용어: 가설검정, p-값, 신뢰구간, 붓스트랩 방법, 재현성.

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