한국심리학회지: 일반
Korean Journal of Psychology: General
2026, Vol. 45, No. 1, 31-54
http://dx.doi.org/10.22257/kjp.2026.3.45.1.31

베이즈 인자를 이용한 독립표본 t 검정: 등분산 및 이분산 가정에 따른 비교

강 채 린
성균관대학교 심리학과 / 석사

장 승 민†
성균관대학교 심리학과 / 교수

본 연구는 두 집단 분포의 등분산을 가정하는 JZS(Jeffreys-Zellner-Siow) 사전분포 설정과 이분산을 허용하는 BFGC(Girón-del Castillo) 사전분포 설정을 이용한 베이지안 독립표본 t 검정의 타당성을 비교하였다. 특히 표본크기 비율과 분산 비율의 조합이 베이즈 인자의 산출에 미치는 영향을 시뮬레이션 연구를 통해 확인하였다. 시뮬레이션은 분산 비율(1:1, 2:1), 표본크기 비율(1:1, 2:3, 3:2), 표준화 효과 크기(0, 0.2, 0.5, 0.8), 총 표본크기(50, 100, 200)를 교차한 조건에서 수행되었으며, 각 조건당 500회의 반복을 통해 생성된 총 30,000개의 데이터 세트가 사용되었다. 분석 결과 두 사전분포 설정이 산출한 베이즈 인자는 이분산 조건에서 분산이 큰 집단의 표본크기의 상대적 크기에 따라 상반된 양상이 확인되었다. JZS 설정은 분산이 큰 집단의 표본크기가 상대적으로 큰 경우 참인 가설에 대한 지지가 약화되는 반면, BFGC 설정은 분산이 큰 집단의 표본크기가 상대적으로 작은 경우 유사한 양상을 보였다. 이러한 결과는 베이지안 t 검정에서 등분산 가정이 표본크기 및 분산 비율과 상호작용하여 베이즈 인자의 크기에 체계적인 영향을 미친다는 점을 보여준다. 또한 이분산이 의심되는 실제 연구 상황에서 등분산 가정에 기반한 JZS 모형보다 이분산을 허용하는 BFGC 모형을 사용하는 것이 더 나은 베이지안 추론을 제공할 수 있음을 시사한다.

주요어 : 베이즈 인자, 베이지안 t 검정, 이분산성, 불균형 표본크기

† 교신저자: 장승민, 성균관대학교 심리학과, 서울특별시 종로구 성균관로 25-2, Tel: 02-760-0692
E-mail: jahngs@skku.edu

독립표본 t 검정은 행동과학을 비롯한 경험과학 분야에서 가장 널리 사용되는 통계적 절차 중 하나이다. 이 검정은 두 독립 집단 간 평균 차이가 0이라는 귀무가설(null hypothesis, 영가설)을 검토하며, 통상적으로 두 집단의 모집단 분산이 동일하다는 등분산 가정을 전제로 수행된다. 그러나 실제 연구에서는 등분산 가정이 충족되지 않는 자료를 접하는 경우가 적지 않다(Bryk & Raudenbush, 1988). 잘못된 등분산 가정을 전제로 t 검정을 수행하면, 특히 두 집단의 표본크기가 상이할 때 제1종 오류율을 적절히 통제할 수 없다(Delacre et al., 2017; Welch, 1938). Welch의 t 검정은 빈도주의 통계에서 이러한 문제를 해결하기 위해 제안된 대표적인 방법이다. 최근에는 등분산 가정을 전제로 하는 Student t 검정보다 Welch t 검정을 독립표본 t 검정의 기본 절차로 채택해야 한다는 견해가 점차 지지를 얻고 있다(Delacre et al., 2017; Zimmerman, 2004).

전통적인 가설검정에서는 유의확률(p 값)을 이용하여 귀무가설에 대한 기각 여부를 판단한다. 유의확률은 귀무가설이 참이라는 전제하에서 관측된 결과(데이터) 및 더 극단적인 결과를 얻을 확률을 의미하는데, 이 값이 사전에 설정된 유의수준보다 작으면 귀무가설을 기각하고 그렇지 않으면 기각하지 않는다. 그러나 유의확률은 대부분의 연구자가 데이터를 통해 실제로 알고자 하는 '대립가설(alternative hypothesis, 상대가설)이 참일 확률' 또는 '귀무가설이 참일 확률'을 알려주지 않는다. 또한 '대립가설이 귀무가설에 비해 얼마나 더 그럴듯한가'에 대해서도 답해 주지 않는다.

유의확률을 이용한 가설검정의 한계(Berger & Delampady, 1987; Cohen, 1994)는 오래전부터 지적되었는데, 이에 대한 대안의 하나로 베이즈 인자(Bayes Factor, BF)를 이용한 베이지안 가설검정이 주목을 받아 왔다(Hoijtink et al., 2019; Wagenmakers et al., 2018). 베이즈 인자란 관측된 데이터에 대한 두 가설의 주변우도(marginal likelihood) 비율로서, 데이터가 한 가설을 다른 가설에 비해 얼마나 더 지지하는지를 정량화한다. 예를 들어 BF_{10}이 10이라면 데이터가 대립가설(H_1)을 귀무가설(H_0)보다 10배 더 지지한다는 의미이며, 0.1이라면 반대로 귀무가설을 10배 더 지지한다는 뜻이다. 따라서 베이즈 인자는 p 값처럼 '귀무가설 하에서 관측 데이터를 얻을 가능성이 얼마나 낮은가'가 아니라, 관측 데이터를 바탕으로 '어느 가설이 얼마나 더 그럴듯한가'를 직접적으로 판단할 수 있는 정보를 제공한다(Gönen et al., 2018; Wagenmakers et al., 2016). 베이지안 t 검정에서는 효과 크기가 0이라는 가설(예, $\mu_1 - \mu_2 = 0$)과 그렇지 않다는 가설(예, $\mu_1 - \mu_2 \neq 0$) 간의 상대적 신빙성을 베이즈 인자를 통해 비교할 수 있다.

베이즈 인자 산출에 필요한 주변우도를 계산하기 위해서는 추정 모수에 대해 적절한 사전분포를 설정해야 한다. 분산분석이나 회귀분석 같은 선형모형에서는 효과 크기에 대한 사전분포로 Jeffreys-Zellner-Siow(JZS) 설정이 널리 사용된다(Schmalz, et al., 2023). 이러한 경향은 *Psychological Science*와 같은 일반 학술지에서 확인할 수 있다. 이 학술지에 2024년에서 2025년 사이에 게재된 논문 중 베이즈 인자를 보고한 논문은 19편이었고 그중 일반선형모형에 관한 것은 15편이었다. 이 중 사전분포가 명시되지 않은 3편을 제외한 12편에서 모두 JZS 사전분포가 사용되었다.

베이지안 t 검정에서 JZS 설정은 효과 크기의 사전분포로 코쉬(Cauchy) 분포를 사용한다.

이 설정은 효과 크기에 대한 사전 지식을 강하게 요구하지 않으면서도 비현실적으로 큰 효과 크기에 낮은 가능성을 부여함으로써 다양한 연구 맥락에서 객관적이고 합리적인 추론을 가능하게 한다(Rouder et al., 2009). 아울러 마르코프 체인 몬테카를로(MCMC) 샘플링과 같은 복잡한 절차 없이 베이즈 인자를 산출할 수 있다는 실용적 이점도 있다.

그러나 JZS 사전분포를 사용하는 베이지안 독립표본 t 검정은 두 집단 분포의 등분산 가정에 기초하고 있다. 앞서 살펴본 바와 같이 빈도주의 접근에서는 독립표본 t 검정에서 등분산 가정의 적절성 여부가 중요하게 다루어진다. 그러나 베이지안 접근에서는 등분산 가정이 충족되지 않을 때 이것이 베이즈 인자의 산출과 해석에 어떠한 영향을 미치는지에 대해 알려진 바가 많지 않다. 따라서 JZS 설정하에 수행된 베이지안 t 검정이 이분산 조건에서 어떻게 작동하는지에 대한 체계적인 검토가 필요하다. 또한 등분산을 가정하지 않고 JZS 설정과 같은 방식으로 베이즈 인자를 구하는 절차를 확인하고 비교할 필요도 있다.

본 연구는 베이지안 독립표본 t 검정에서 등분산을 가정하는 사전분포 설정과 이분산을 허용하는 사전분포 설정에서 산출된 베이즈 인자의 특성을 체계적으로 검토하고자 하였다. 구체적으로는 JZS 설정과 대안적 설정을 이용한 베이지안 t 검정이 각각 제1종 오류와 제2종 오류에서 어떠한 차이를 보이는지, 그리고 이러한 차이가 표본크기 비율과 분산 비율에 따라 어떻게 달라지는지를 확인하였다. 이를 통해 실제 연구에서 표본크기 비율과 분산 비율에 따라 적절한 베이지안 모형을 선택할 수 있는 근거를 제시하고자 하였다.

독립표본 t 검정과 이분산성

독립표본 t 검정은 두 집단의 평균 차이에 대한 가설을 검정하기 위해 일반적으로 관측값 분포의 독립성, 정규성, 등분산성을 가정한다. 빈도주의 통계에서는 등분산 가정하에서 두 집단의 분산 추정치 s_1^2과 s_2^2을 합동 분산 추정치 s_p^2으로 통합한다.

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

합동 분산 추정치를 이용한 Student의 t 통계량은 자유도가 n_1+n_2-2인 t 분포를 따른다.

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

두 모집단 분산이 실제로는 다른데도 등분산을 가정한 t 검정을 사용하면, 특히 두 집단의 표본크기가 다를 때, 검정 통계량의 분포가 왜곡되고 제1종 오류율이 정확히 통제되지 못한다. 즉, 귀무가설이 참임에도 이를 기각할 확률이 유의수준을 초과하거나 미달한다.

독립적인 두 정규 모집단의 평균 차이를 검정하거나 신뢰구간을 추정할 때, 미지의 모집단 분산이 서로 다르다고 가정($\sigma_1^2 \neq \sigma_2^2$)하는 상황을 통계학에서는 베렌스-피셔 문제(Behrens-Fisher problem)라고 부른다(Dayal & Dickey, 1976; Giron & Castillo, 2021; Kim & Cohen, 1998). 모집단 분산이 다를 경우 합동

분산을 사용하지 않는 t 통계량은 다음과 같이 표현된다.

$$t_W = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

베렌스-피셔 문제의 핵심은 t_W가 따르는 확률분포(베렌스-피셔 분포)를 정확히 특정하기 어렵다는 점에 있다. Welch는 Satterthwaite의 근사식으로 추정된 유효 자유도(effective degrees of freedom)를 가지는 Student의 t 분포가 베렌스-피셔 분포의 근사 분포로 사용될 수 있음을 보였다(Satterthwaite, 1946; Welch 1938, 1947). Welch의 방식은 베렌스-피셔 문제에 대해 제안된 빈도주의적 해법 중에서 가장 널리 받아들여지며 이를 이용한 절차가 Welch의 t 검정이다.

빈도주의와 베이지안 가설검정

전통적인 빈도주의 통계학에서 가설검정은 귀무가설 하에서 관측된 데이터(또는 더 극단적인 데이터)가 나타날 확률(p 값)에 의존한다. 이 확률을 유의수준(예, .05)과 비교하여 귀무가설의 타당성에 대한 이분법적인 의사결정(기각 또는 기각 실패)을 내리는데 이러한 관행은 오랫동안 비판의 대상이 되어 왔다. 가장 큰 문제는 p 값이 연구 가설의 참/거짓에 대한 확률적 정보를 제공하지 않는다는 점이다. 유의확률은 '가설(H)하에서 데이터(D)가 관측될 확률', 즉 $P(D \mid H)$ 을 알려준다. 그러나 대부분의 연구자는 '데이터가 주어졌을 때 가설이 참일 확률'인 $P(H \mid D)$ 에 더 관심이 많다.

또한 p 값은 표본크기에 과도하게 민감하여, 표본이 충분히 크면 실질적인 의미가 없는 미세한 차이조차 통계적으로 유의하다고 판정하는 경향이 있다(Berger & Delampady, 1987; Cohen, 1994). 더욱이 빈도주의적 접근법은 귀무가설(예: 두 집단의 평균이 같다)을 기각할 증거를 찾을 뿐 귀무가설이 참이라는 증거를 직접적으로 제공하지 못하며, 심지어 귀무가설에 불리한 증거를 과장하는 경향이 있다(Lakens et al., 2018; Rouder et al., 2009).

유의확률과 달리 베이즈 인자는 귀무가설과 대립가설이 데이터에 의해 지지받는 상대적인 증거의 강도를 정량화한다(김달호, 2013, pp. 159-165). 베이즈 인자 BF_{10}은 귀무가설 대비 대립가설에 대한 데이터의 평균적 우도 비율로, BF_{10}이 1보다 크면 데이터가 대립가설에 대한 믿음을 더 높인 것을, 1보다 작으면 귀무가설에 대한 믿음을 더 높인 것을 의미한다. 일반적으로 BF_{10}이 1에서 3사이면 대립가설을 지지하는 참고 수준의 증거, 3에서 10 사이면 약한 증거, 10에서 100 사이면 강한 증거, 100 이상이면 아주 강한 증거로 해석한다(Jeffreys, 1961).[1] 이처럼 베이즈 인자는 유의확률을 이용하는 전통적인 접근과 달리 귀무가설 대비 대립가설을 지지하는 증거의 정도를 연속적인 수치로 정량화할 수 있으며, 자료수집 계획(Stopping rule)에 영향을 받지 않고 증거를 해석할 수 있다는 점에서 통계적 추론의 일관성을 제공한다(Berger & Delampady, 1987; Gönen et al., 2018; Weltzels et al., 2011).

1) BF_{01}은 대립가설 대비 귀무가설에 대한 데이터의 주변 우도 비율이며 수치의 의미는 BF_{10}에 대한 역수로 해석한다.

베이즈 인자를 계산하기 위해서는 각 가설 하에서 모수에 대해 사전분포를 설정해야 하는데, 사전분포를 다르게 설정하면 주변우도와 베이즈 인자의 값도 달라진다. 따라서 명확한 사전 정보가 존재하지 않는 상황이라면, 데이터의 척도나 단위에 영향을 받지 않으면서 연구자의 편향이나 주관성의 개입을 최소화하는 객관적 사전분포를 설정하는 것이 바람직하다(Berger & Pericchi, 1996). 독립표본 t 검정의 맥락에서는 앞서 언급한 JZS(Jeffreys-Zellner-Siow) 설정이 객관적 사전분포를 사용하는 베이지안 가설검정의 표준적인 방법으로 자리 잡았다.

등분산 조건에서의 베이지안 t 검정

독립표본 t 검정에서 빈도주의와 베이지안 접근의 근본적인 차이는 효과 크기 ($\Delta = \mu_1 - \mu_2$)를 바라보는 관점에 있다. 빈도주의 관점에서는 효과 크기를 고정되어 있으나 알 수 없는 상수로 취급한다. 반면 베이지안 관점에서는 효과 크기를 불확실성을 가진 변수로 취급하여 확률분포로 표현한다. 따라서 베이지안 접근에서는 각 가설(귀무가설과 대립가설)하에서 효과 크기에 대한 사전분포를 명시적으로 설정한다. 이후 관측된 데이터를 통해 사전분포를 사후분포로 갱신하고, 베이즈 인자를 산출하여 가설 간 상대적 타당성을 평가한다(Gelman, et al, 2013).

베이즈 인자를 산출하기 위해서는 일반적으로 적절한 사전분포를 선택해야 하고 복잡한 적분 계산이 수반된다(Wagenmakers et al., 2010). 특정 모수에 대한 점 귀무가설(예, $\Delta = 0$)에 대해서는 사전분포와 사후분포의 밀도비를 이용하여 베이즈 인자를 계산하는 Savage-Dickey ratio method(Dickey, 1971, 이하 SD 방법)가 사용될 수 있다. 이 방법을 적용하려면 사후분포의 밀도를 추정해야 하는데 대개 MCMC 샘플링과 같은 절차가 필요하다. SD 방법의 대안으로 비교적 간단한 통계 모형에서 특정한 사전분포를 사용하는 경우에는 샘플링 없이도 해석적 방법으로 주변우도를 계산하여 베이즈 인자를 쉽게 계산할 수 있다.

베이지안 t 검정에서는 집단 j의 관측값 X_j가 평균이 μ_j, 표준편차가 σ_j인 정규분포를 따른다고 가정하고, 귀무가설과 대립가설에 대해 서로 다른 사전 가정을 부여하여 두 가설을 비교한다. 이때 귀무가설은 효과 크기가 0이라는 점가설로, 대립가설은 효과 크기가 특정 사전분포를 따른다는 분포가설로 설정된다. 각 모수에 대한 사전분포는 전통적으로 Jeffreys(1937)가 제안한 무정보 사전분포(이하 제프리스 사전분포)를 사용해왔다. 이때 위치 모수(평균 또는 평균 차이)에 대해서는 균등분포를, 척도 모수(각 집단의 분산)에 대해서는 역수에 비례하는 부적절 사전분포를 사용한다.

Jeffreys 이후에도 무정보 사전분포에 대한 여러 제안이 있었는데 심리학 분야에서는 효과 크기 $\Delta = \mu_1 - \mu_2$에 대해 Rouder et al. (2009)이 종합하여 제안한 계층적 사전분포 구조가 가장 널리 사용된다. 이 계층적 구조에서 Δ는 평균이 0이고 분산이 σ_Δ^2인 정규분포를 따르며, 효과 크기의 분산 σ_Δ^2는 척도 역카이제곱 분포를 따른다(Zellner & Siow, 1980).

$$\Delta \sim N(0, \sigma_\Delta)$$
$$\sigma_\Delta^2 \sim Scale-inv-\chi^2(1, r^2)$$

이 계층적 구조를 결합하면 $\delta = \Delta/\sigma_{\Delta}$에 대한 코쉬 사전분포 설정으로 단순화된다. 코쉬 분포는 꼬리가 두꺼운 특성을 가져 데이터가 강한 증거를 보일 때 사전분포의 영향력을 적절히 희석시키며, 연구자의 주관적 믿음을 배제하고 객관성을 확보하기에 유리하다. 이때 두 집단의 오차 분산은 같다고 가정하고 ($\sigma_j^2 = \sigma^2$) 사전분포로는 제프리스 사전분포를 적용한다(Jeffreys, 1961). 효과 크기와 오차 분산에 대한 이와 같은 사전분포 조합을 JZS 사전분포라고 한다.

이러한 설정은 베이즈 인자를 구하기 위한 계산적 측면에서도 중요한 장점을 제공한다. 베이지안 t 검정에서 JZS 사전분포를 사용하면 t 통계량과 표본크기(n_1, n_2)만으로도 수치적분을 이용하여 베이즈 인자를 얻을 수 있다. 이러한 장점 덕분에 R의 BayesFactor 패키지(Morey et al., 2018), IBM SPSS(IBM Corp., 2023), jamovi(The jamovi project, 2025)와 JASP(JASP Team, 2025)의 Bayesian 모듈과 같은 주요 통계 소프트웨어에서는 JZS 사전분포를 t 검정을 포함한 베이지안 선형모형 분석의 표준으로 채택하고 있다(Van Ravenzwaaij & Etz, 2021).

그러나 JZS 사전분포의 베이즈 인자의 계산에는 등분산 가정하에서 유도된 식이 사용된다. 연구자들은 JZS 사전분포의 편리함에 의존하지만, 실제는 이분산인 자료를 JZS 사전분포 설정으로 분석하는 것이 얼마나 타당한가에 대해서는 알려진 바가 많지 않다.

이분산 조건에서의 베이지안 t 검정

JZS 기반 베이지안 t 검정은 등분산 가정을 전제로 한다는 한계가 있다. 그렇다면 이러한 가정 없이 베이즈 인자를 어떻게 산출할 수 있는가? Jeffreys(1937, 1940)는 베이지안 t 검정을 제안한 초기에 이미 각 집단의 분산에 독립적인 사전분포를 적용하여 이분산을 허용하는 접근을 제안하였다. 그러나 제프리스 사전분포가 부적절 사전분포이기 때문에, 독립된 두 개의 분산 사전분포를 설정하여 단일 효과크기 모수에 대한 사후분포를 도출하는 과정에 추가적인 계산적 어려움이 발생하였고 이를 극복하기 위해 이후 다양한 방법이 제안되었다(Box & Tiao, 1973; Dayal & Dickey, 1976; Berger & Delampady, 1987; Moreno et al., 1999; Wetzels et al., 2009).

베이지안 t 검정에서 두 집단의 개별 분산을 다루면서 단일 베이즈 인자를 산출하는 한 가지 방법은 MCMC 추정과 SD 방법을 활용한다(Wetzels et al., 2009; Kruschke, 2013; Maier et al., 2024). Wetzels et al. (2009)은 각 집단의 분산에 독립적인 사전분포를 설정하고, MCMC로 생성한 사후분포에 SD 방법을 적용하여 효과크기에 대한 베이즈 인자를 산출하였다. Maier et al. (2024)은 이 접근법을 확장하여 모형 평균화 베이지안 t 검정을 제안하였다.

한편 Girón and Del Castillo (2021)는 고유베이즈 인자(intrinsic Bayes factor; Berger & Pericchi, 1996; Gu et al., 2017; Moreno et al., 1999)의 구조적 틀을 바탕으로 제프리스 사전분포의 비정규성 문제를 해결하는 베이지안 이분산 t 검정을 제안했다. 이들은 객관적 사전분포를 설정한 이분산 조건에서 표본평균 차이, 표본분산, 표본크기, 자유도를 이용하여 베이즈 인자를 구하는 방법을 제시하였으며, JZS 사전분포와 마찬가지로 MCMC 샘플링 없

이 수치적분을 통해 베이즈 인자를 구할 수 있도록 하였다. 이 절차에서 유도된 사후 분포는 베렌스-피셔 분포를 따르며, 제안자들의 이름(Girón & del Castillo)을 반영하여 이하에서는 이들의 사전분포 설정 및 베이즈 인자 계산 절차를 BFGC로 지칭한다.

Maier et al. (2024)은 두 집단의 표본크기가 불균형할 때 베이지안 등분산 t 검정과 이분산 t 검정의 베이즈 인자를 비교하였다. 이들은 분산이 큰 집단의 표본크기가 클 때는 이분산 모형이, 반면 그 표본크기가 작을 때는 등분산 모형이 대립가설에 대해 더 강한 증거를 제공하는 경향이 있음을 발견하였다. 이는 표본크기와 분산의 상호작용이 베이지안 t 검정의 추론 결과에 실질적인 영향을 미친다는 것을 보여준다. 그러나 이 연구에서는 이러한 상호작용이 왜 발생하는지, 그리고 어떤 조건에서 문제가 두드러지는지에 대해 체계적인 설명이 제시되지 않았다.

이와 유사하게 등분산 사전분포와 이분산 사전분포를 갖는 베이지안 모형을 비교한 다른 연구들도 다양한 표본크기와 분산 조합을 체계적으로 검토하지 못한 경우가 많았다. 특히 여러 연구에서 Box and Tiao(1973)의 예제 데이터를 활용하였으나, 이 데이터만으로 표본크기 불균형과 분산의 관계를 포괄적으로 탐색하기에는 한계가 있다(Girón & Del Castillo, 2021; Moreno et al., 1999; Wetzels et al., 2009).

본 연구는 이분산 조건에서 베이지안 t 검정의 등분산 가정이 언제, 어떻게 문제가 되는지를 탐색하였다. 이를 위해 시뮬레이션을 통해 효과 크기, 표본크기 비율, 분산 비율의 조합을 달리한 조건에서 JZS 사전분포 기반 등분산 모형의 결과와 BFGC 사전분포 기반 이분산 모형의 결과를 비교함으로써, 등분산 가정이 베이즈 인자 산출에 미치는 영향을 체계적으로 규명하고자 하였다.

방 법

시뮬레이션 설계 및 자료 생성

시뮬레이션 설계 조건은 분산 비율과 표본크기 비율의 조합, 표준화 효과 크기(코헨의 d), 그리고 총 표본크기 세 가지였다(표 1).

분산 비율은 두 집단의 분산비(var_1 : var_2)가 1 대 1로 등분산인 조건과 2 대 1로 이분산인 조건 두 가지를 포함하였다. 다음으로 표본크기 비율은 균등 표본크기와 불균등 표본크기 조건을 설정하였다. 불균등 표본크기의 비율(n_1 : n_2)은 2대 3 또는 3대 2로 설정하였으며 이는 Delacre et al. (2017)의 대표 조건에서 차용하였다.

분산 비율과 표본크기 비율의 조합은 다섯 가지로 구성하였다. 등분산 조건에서는 균등 표본크기 조건(A)과 불균등 표본크기 조건(B)을 설정하였다. 이분산 조건(var1 : var2 = 2:1)에서는 균등 표본크기 조건(C), 분산이 큰 집단의 표본크기가 더 작은 조건(D, n_1 : n_2 = 2 : 3), 그리고 분산이 큰 집단의 표본크기가 더 큰 조건(E, n_1 : n_2 = 3 : 2)을 설정하였다.

표준화 효과 크기 d는 0, 0.2, 0.5, 0.8의 네 수준으로 설정하였다(Cohen, 1988). 베이즈 인자는 귀무가설이 참인 경우에는 귀무가설에 대한 데이터의 지지 정도를 보여준다. 따라서 효과 크기가 0일 때는 베이즈 인자가 귀무가설을 선호하는 양상을 확인하였다. 0이 아닌 효과 크기에 대해서는 크기가 커지면 베이즈 인자가 대립가설을 더 강하게 선호하는지 확

표 1. 시뮬레이션 비교 조건

비율 조합	분산비(var1 : var2)*	표본크기비	표준화 효과 크기(d)	총 표본크기(N)
A	등분산(1:1)	균등(1:1)	0, 0.2, 0.5, 0.8	50, 100, 200
B	등분산(1:1)	불균등(2:3)		
C	이분산(2:1)	균등(1:1)		
D	이분산(2:1)	불균등(2:3)		
E	이분산(2:1)	불균등(3:2)		

*var1은 집단1의 분산, var2는 집단2의 분산을 의미

인하였다.

두 집단의 표본크기 합인 총 표본크기는 50, 100, 200의 세 수준으로 설정하였다. 베이즈 인자에서 표본크기는 증거 강도로 기능하므로, JZS와 BFGC 두 방법 모두 표본크기 증가에 따른 일관된 증거 축적 패턴을 보이는지 확인하였다. 동일한 효과 크기 조건에서는 표본크기가 증가할수록 참인 가설에 대한 지지 정도가 높아져야 한다.

데이터는 다음과 같이 생성하였다:

1. 표본크기·분산 비율 조합, 효과 크기, 표본크기 조건을 교차하여 5×4×3의 기본 설정을 생성

2. 효과 크기와 각 집단의 분산을 고정하고 각 집단의 평균값을 산출

가. 집단 1의 분산은 4로 고정하고 집단 2의 분산을 등분산 조건은 4, 이분산 조건은 2로 설정

나. 표준화 효과 크기의 분모인 표준편차는 두 집단 분산의 산술평균의 제곱근으로 정의[2]

다. 집단 1의 평균을 0으로, 집단 평균 차이를 $d \times \sqrt{(var_1 + var_2)/2}$로 설정

3. 60개 조건당 500번 반복하여 총 30,000개의 데이터 세트를 생성

자료의 생성 및 분석은 모두 R 4.3.0(R Core Team, 2023)을 이용하여 수행되었다. 각 데이터 세트에 대해 JZS 사전분포 기반 베이즈 인자(BF_{JZS})와 BFGC 사전분포 기반 베이즈 인자(BF_{BFGC})를 산출하였다. JZS 베이즈 인자는 R의 BayesFactor 패키지(Morey & Rouder, 2024)의 ttestBF 함수를 사용하여 산출하였으며, 코쉬 분포의 스케일(r)은 1로 설정하였다. 사전분포의 정보량이 베이즈 인자의 크기에 영향을 미치므로, 두 방법 간 베이즈 인자를 직접 비교하기 위해서는 유사한 정보량을 갖는 사전분포를 설정할 필요가 있다. 예비분석 결과, JZS 사전분포의 r = 1(wide)일 때 BFGC 사전분포와 전반적인 귀무가설 선호 정도가 가장 유사하였다. BayesFactor 패키지의 ttestBF 함수의 기본값($r = 1/\sqrt{2}$)을 설정한 경우에도 주요 결과의 패턴에서는 차이가 없었다.

BFGC 베이즈 인자는 Girón and Del Castillo (2021)의 식을 참조하여 저자들이 함수를 생성하여 산출하였다. 수치적분은 pracma 패키지

2) 이는 Cohen (1988)의 효과 크기 정의를 따르는 것으로, 등분산 조건에서는 기존 방식과 동일한 결과를 산출하면서도, 이분산 조건에서는 두 분산을 동등하게 고려함으로써 두 분산 가정을 모두 포괄한다.

표 2. 베이즈 인자의 해석 기준(Jeffreys, 1961)

$\log_{10}BF_{10}$	BF_{10}	귀무가설에 반하는 증거
0에서 0.5	1에서 3.2	참고수준의 증거
0.5에서 1	3.2에서 10	실질적인 증거
1에서 2	10에서 100	강한 증거
> 2	> 100	결정적인 증거

(Borchers, 2023)의 integral 함수를 사용하였고 방법은 Gauss-Kronrod를 선택하였다.

모든 베이즈 인자는 밑이 10인 상용로그로 변환하여 분석에 사용하였다. 시뮬레이션 연구에서 베이즈 인자는 극단값에 민감하여, 하나의 데이터 세트 내에서도 소수점 이하부터 백만 이상까지 넓은 범위의 값이 산출될 수 있다. 로그 변환을 사용하면 이러한 극단값의 영향을 줄이면서 분석할 수 있다. BF_{10}과 BF_{01}은 역의 관계이며($BF_{01} = 1/BF_{10}$), 로그 변환 시에는 $\log_{10}BF_{01} = -\log_{10}BF_{10}$의 관계가 성립한다. $\log_{10}BF_{10}$이 양수면 대립가설, 음수면 귀무가설을 선호함을 의미하며, 절댓값 0.5 이상은 실질적 증거, 1 이상은 강한 증거, 2 이상은 결정적 증거로 해석한다(표 2).

분석

본 연구는 두 가지 문제를 순차적으로 검증하였다. 먼저 JZS 사전분포가 다양한 조건에서 일관된 결과를 산출하는지 확인하기 위해 분산 비율과 표본크기 비율의 조합별 베이즈 인자를 살펴보았다. 이어서 각 조합별 BFGC 베이즈 인자를 확인하고, 두 방법의 산출 패턴을 비교하였다.

시뮬레이션 기반의 베이즈 인자를 비교하는 대표적인 표준 지표는 아직 확립되지 않았으나, 몇 가지 해석 지표들이 제안되어 왔다(Gönen et al., 2019; Kelter, 2020, 2021). 본 연구는 Kelter (2020, 2021)가 제안한 임계값 기반 오류율을 비교 지표로 채택하였다. 이 방법은 실제 연구자들이 베이즈 인자를 해석할 때 사용하는 증거의 강도 기준을 시뮬레이션 평가에 적용한 것으로(표 2), 베이즈 인자를 가설검정의 도구로 사용하는 맥락에 부합한다. Kelter(2020)는 $BF_{10} = 3$을 기준값으로 설정하고 각 시뮬레이션 조건이 이를 얼마나 안정적으로 넘어서는지 평가하였다. 본 연구에서는 $BF_{10} = 3$ 대신 $\log_{10}BF_{10} = 0.5$를 사용하여, 효과가 없는 조건(H_0 참)에서 $\log_{10}BF_{10} \geq 0.5$를 산출하는 것을 제1종 오류, 효과가 존재하는 조건(H_1 참)에서 $\log_{10}BF_{10} < 0.5$를 산출하는 것을 제2종 오류로 간주하였다.

이러한 지표는 특정 절차가 주어진 조건에서 얼마나 '신뢰할 만한 증거의 강도'를 산출하는지 비교하는 데 목적이 있다. 단, 여기서의 오류율은 빈도주의적 가설검정의 공식적 오류 개념과 달리 베이즈 인자의 실용적 성능을 평가하기 위한 시뮬레이션 지표이며(Kelter, 2021), 본문에 제시되는 오류율과 그 차이는 각 방법론이 산출하는 베이즈 인자의 분포적 특성에 따른 수치적 차이일 뿐 특정 방법의 우월성을 판단하는 객관적 기준이 아님에 유의해야 한다.

본 연구에서는 로그 베이즈 인자의 평균값, 오류율, 그리고 그 차이를 종합적으로 살펴봄으로써 효과 크기 및 분산 비율·표본크기 비율 조합에 따라 각 방법이 올바른 가설을 얼마나 적절한 강도로 지지하는지를 비교하였다. 추가로 각 방법 내에서 이분산성과 표본크기 비율이 베이즈 인자 산출에 미치는 영향을 체

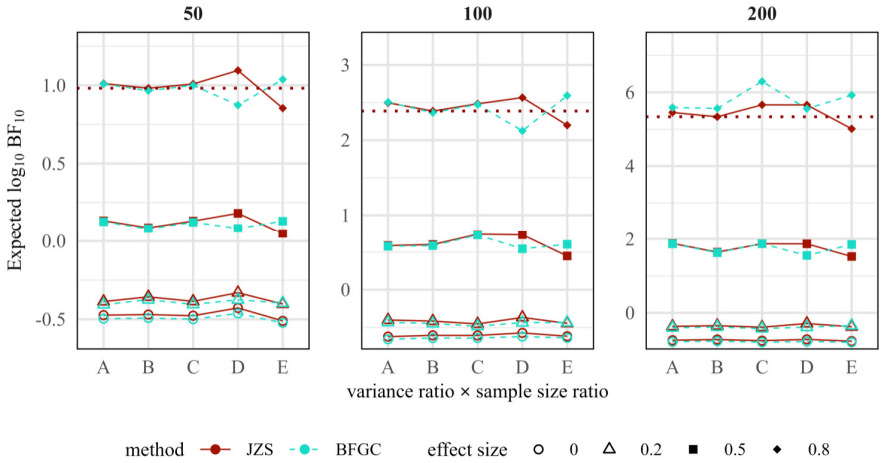

그림 1. JZS와 BFGC 사전분포 설정의 조건별 로그 베이즈 인자의 평균

계적으로 분석하기 위해, 다섯 가지 조합에 대한 네 가지 대비를 설정하여 오류율의 차이를 비교하였다:

(대비1) 조합 A, B와 조합 C, D, E: 등분산 조건과 이분산 조건

(대비2) 조합 A와 B: 등분산 조건에서 표본크기 비율이 1이거나 1이 아닌 조합

(대비3) 조합 C와 조합 D, E: 이분산 조건에서 표본크기 비율이 1이거나 1이 아닌 조합

(대비4) 조합 D와 조합 E: 이분산, 불균등 표본크기 조건에서 분산이 큰 집단의 표본크기가 큰 조합과 분산이 큰 집단이 표본크기가 작은 조합

이를 통해 개별 조합 간 평균 차이뿐 아니라, 분산 동등성(등분산 vs. 이분산), 표본크기 균등성(균등 vs. 불균등), 그리고 분산비·표본크기비 관계(분산이 큰 집단의 표본크기가 큰 경우 vs. 작은 경우)가 베이즈 인자에 미치는 영향을 체계적으로 파악하고자 하였다.

결 과

요약통계량 비교

표 3은 시뮬레이션 요인의 수준에 따른 JZS와 BFGC 방법의 로그 베이즈 인자에 대한 요약통계량이다. 귀무가설이 참인 조건(d = 0)에서는 JZS와 BFGC 방법 모두 $\log_{10}BF_{10} < 0$으로 영가설을 선호하였으며(BF_{JZS}: -0.605, BF_{BFGC}: -0.639), 조합 간 차이는 미미하였다. 전반적으로 BFGC 방법이 JZS 방법보다 영가설을 약간 더 선호하는 경향을 보였다. 대립가설이 참인 조건(d = 0.2, 0.5, 0.8)에서는 두 방법 모두 효과 크기가 증가할수록 $\log_{10}BF_{10}$이 증가하여 대립가설을 지지하는 일관된 패턴을 보였다. JZS 방법의 경우 효과 크기 0.2일 때 -0.381, 0.5일 때 0.809, 0.8일 때 2.946으로 증가하였다. 표본크기에 따라서도 N = 50일 때 0.065, N = 100일 때 0.501, N = 200일 때 1.511로 증가하였다. 이러한 패턴은 BFGC 방법에서도 유사하게 나타났다.

효과 크기가 0.2인 경우, 두 방법 모두 평균

표 3. JZS와 BFGC 사전분포 설정의 조건별 로그 베이즈 인자($\log_{10}BF_{10}$)의 평균 및 표준편차

요인	수준	JZS		BFGC		평균 차이	$\frac{BF_{JZS}}{BF_{BFGC}}$
		평균	표준편차	평균	표준편차		
분산비 × 표본크기비	A	0.708	1.810	0.702	1.850	0.006	1.014
	B	0.687	1.770	0.685	1.830	0.002	1.005
	C	0.717	1.860	0.751	2.030	-0.034	0.925
	D	0.777	1.860	0.636	1.800	0.141	1.384
	E	0.572	1.675	0.740	1.947	-0.167	0.681
효과 크기	0	-0.605	0.116	-0.639	0.123	-0.034	1.081
	0.2	-0.381	0.042	-0.409	0.035	0.028	0.938
	0.5	0.809	0.701	0.802	0.702	-0.007	1.016
	0.8	2.946	1.919	3.058	2.095	-0.111	0.774
표본크기	50	0.065	0.596	0.050	0.598	0.016	1.038
	100	0.501	1.236	0.477	1.245	0.024	1.057
	200	1.511	2.511	1.582	2.678	-0.071	0.849

베이즈 인자가 음수로 나타나 대립가설에 대한 실질적 증거($\log_{10}BF_{10} \geq 0.5$)를 제공하지 못하였다. 이러한 양상은 모든 표본크기와 조합 조건에서 일관되게 나타났다(그림 1). 등분산 조건과 이분산-균등 표본 조건(조합 A, B, C)에서 두 방법 간 평균 차이는 0.1 미만이었다. 조합 별로 살펴보면, JZS 방법의 경우 조합 A부터 D까지 평균이 0.687 ~ 0.777 범위였으며, BFGC 방법은 조합 A, B, C, E가 이와 유사한 범위(0.685 ~ 0.751)를 보였다.

주목할 점은 이분산, 불균등 표본크기 조건(조합 D, E)에서 두 방법 간 상반된 패턴이 나타났다는 것이다(그림 1). 이러한 교차 패턴은 효과 크기가 0.5 이상일 때 뚜렷하게 나타났으며, 표본크기가 증가해도 유지되었다. JZS 방법은 분산이 큰 집단의 표본크기가 작은 조합 D(평균 0.777)가 분산이 큰 집단의 표본크기가 큰 조합 E(평균 0.572)보다 약 0.20 높은 베이즈 인자를 산출하였다. 예를 들어, d = 0.5, N = 100인 조건에서 조합 D와 조합 E의 BF_{JZS} 차이는 0.203이었다. 이는 원척도에서 약 1.6배의 베이즈 인자 차이를 의미한다 ($10^{0.203} \approx 1.596$). 반면 BFGC 방법은 조합 E(평균 0.740)가 조합 D(평균 0.636)보다 약 0.10 높은 베이즈 인자를 산출하였다. 즉, JZS 방법과 정반대의 패턴을 보였다. 이러한 조합 D와 조합 E의 두 방법의 베이즈 인자가 역전되는 패턴이 N = 100, 200이고 효과 크기 0.5 이상인 조건에서 명확하게 나타난다.

그림 1은 효과 크기별로 표본크기와 조합에 따른 두 방법의 평균 로그 베이즈 인자를 나타낸 것이다. 그림 1에서 확인할 수 있듯이, 효과 크기 0.2 조건(속이 빈 원)은 모든 조합에서 베이즈 인자가 음수를 유지하여 다른 효

과 크기 조건들과 뚜렷이 구분된다.

임계값 기반 오류율 분석

제1종 오류율 (d = 0)

두 방법 모두 모든 조건에서 제1종 오류율이 .03 이하로 매우 낮았다(표 4). 방법 간 차이는 1%p 미만이었다. 이는 로그 베이즈 인자의 요약통계량 결과(표 3, 그림 1)에서 본 바와 같이, 두 방법 모두 귀무가설이 참일 때 잘못된 증거($\log_{10}BF_{10} \geq 0.5$)를 제시할 가능성이 극히 낮음을 의미한다.

제2종 오류율 (d = 0.2, 0.5, 0.8)

제2종 오류율 패턴도 표 2 및 그림 1의 결과와 일치하였다(표 5). 효과 크기가 작을 때(d = 0.2) 두 방법 모두 높은 제2종 오류율을 보였다(90% 이상). 특히 표본크기가 작은 경우(N = 50) 오류율은 90~96%에 달했으며, 방법 간 차이는 2%p 이하로 미미하였다. 효과 크기와 표본크기가 증가할수록 오류율은 두 방법 모두에서 뚜렷하게 감소하였다.

등분산 조건과 이분산 조건 간 오류율 차이(대비1)는 두 방법 모두 1~2%p로 매우 낮았다.

표본크기 불균등의 영향

등분산 조건에서 표본크기 불균등(조합 A vs B, 대비 2)의 영향은 대체로 미미하였다. JZS와 BFGC 방법 모두 대부분의 조건에서 2%p 이하의 차이를 보였다. 그러나 특정 조건(d = 0.5, N = 200)에서는 불균등 표본에서 약 4%p 높은 오류율이 나타났다(JZS: 조합 A 18.2% vs 조합 B 22.8%, BFGC: 조합 A 19.2% vs 조합 B 23.4%).

이분산 조건에서는 표본크기 불균등의 영향이 뚜렷했다(대비3). JZS 방법의 경우 d = 0.5, N = 100 조건에서 균등 표본(조합 C: 48.6%)에 비해 불균등 표본(조합 D, E 평균: 약 55%)의 오류율이 약 6.4%p 높았다. BFGC 방법도 유사한 패턴을 보였다. 효과 크기와 표본크기가 증가해도 불균등 표본에서 2%p가량 일관되게 높은 오류율이 유지되었다. 이는 두 방법 모두 이분산 조건에서 표본크기 불균형에 더 영향을 받음을 시사한다.

분산·표본크기 교차 패턴 (대비 4)

가장 주목할 만한 발견은 이분산, 불균등

표 4. JZS와 BFGC 사전분포 설정의 조건별 제1종 오류율($P(\log_{10}BF_{10} \geq 0.5)$)

효과 크기	표본 크기	방법	분산비 × 표본크기비					평균 비교 대비			
			A	B	C	D	E	1	2	3	4
0	50	JZS	.012	.018	.012	.026	.010	-.001	-.006	-.006	.016
		BFGC	.012	.020	.012	.022	.010	.001	-.008	-.004	.012
	100	JZS	.000	.014	.010	.022	.020	-.010	-.014	-.011	.002
		BFGC	.000	.012	.010	.020	.024	-.012	-.012	-.012	-.004
	200	JZS	.006	.014	.002	.018	.004	.002	-.008	-.009	.014
		BFGC	.006	.014	.002	.008	.006	.005	-.008	-.005	.002

표 5. JZS와 BFGC 사전분포 설정의 조건별 제2종 오류율($P(\log_{10}BF_{10} < 0.5)$)

효과 크기	표본 크기	방법	분산비 × 표본크기비					평균 비교 대비			
			A	B	C	D	E	1	2	3	4
0.2	50	JZS	.966	.952	.950	.930	.976	.007	.014	-.003	-.046
		BFGC	.964	.950	.950	.948	.960	.004	.014	-.004	-.012
	100	JZS	.936	.940	.962	.930	.954	-.011	-.004	.020	-.024
		BFGC	.938	.948	.962	.942	.924	.000	-.010	.029	.018
	200	JZS	.922	.912	.914	.862	.914	.020	.010	.026	-.052
		BFGC	.924	.916	.916	.898	.888	.019	.008	.023	.010
0.5	50	JZS	.750	.762	.766	**.736**	**.792**	-.009	-.012	.002	-.056
		BFGC	.748	.766	.766	**.772**	**.746**	-.004	-.018	.007	.026
	100	JZS	.542	.536	.486	**.504**	**.596**	.010	.006	-.064	-.092
		BFGC	.544	.536	.486	**.574**	**.524**	.012	.008	-.063	.050
	200	JZS	.182	.228	.192	**.198**	**.238**	-.004	-.046	-.026	-.040
		BFGC	.192	.234	.198	**.248**	**.198**	-.002	-.042	-.025	.050
0.8	50	JZS	.384	.386	.364	**.346**	**.428**	.006	-.002	-.023	-.082
		BFGC	.382	.376	.360	**.400**	**.368**	.003	.006	-.024	.032
	100	JZS	.084	.074	.070	**.088**	**.094**	-.005	.010	-.021	-.006
		BFGC	.084	.076	.072	**.112**	**.080**	-.008	.008	-.024	.032
	200	JZS	.002	.002	.000	.000	.000	.002	.000	.000	.000
		BFGC	.002	.002	.000	.000	.000	.002	.000	.000	.000

표본 조건 내에서 분산·표본크기 관계에 따라 두 방법의 상대적 성능이 역전되었다는 점이다. 이 패턴은 효과 크기가 0.5 이상일 때 뚜렷하게 나타났으며, 요약통계량에서 관찰된 패턴과 일치하였다(그림 1, 보충 자료 그림 S.1. 참조).

JZS 방법의 경우 분산이 큰 집단의 표본이 더 작은 조합 D에서 일관되게 낮은 제2종 오류율을 보였다. 예를 들어, d = 0.5, N = 100 조건에서 조합 D는 50.4%, 조합 E는 59.6%로 9.2%p 차이가 나타났다. 이는 N = 200 조건에서도 유지되었다(조합 D: 19.8%, 조합 E: 23.8%, 차이 4%p). 효과 크기가 0.8로 증가해도 유사한 패턴이 관찰되었다(N = 100: 조합 D 8.8% vs 조합 E 9.4%).

반면 BFGC 방법의 경우 JZS와 정반대의 패턴을 보였다. 분산이 큰 집단의 표본이 더 큰 조합 E에서 더 낮은 오류율을 나타냈다. d = 0.5, N = 100 조건에서 조합 D는 57.4%, 조합 E는 52.4%로 5.0%p 차이가 나타났다. N = 200 조건에서도 유사한 패턴이 유지되었다(조합 D: 24.8%, 조합 E: 19.8%, 차이 5.0%p).

이러한 역전 효과(3~9%p)는 등분산 조건 내 차이(1~2%p)나 일반적인 표본크기 불균등 효과(2~4%p)보다 크다. 이 패턴은 평균 베이즈 인자 분석(조합 D와 E의 평균 차이 약 0.20)에서도 동일하게 확인되었다. 즉, 두 방법의 성능 차이는 평균적으로도 확인되고, 실질적 증거($\log_{10}BF_{10} \geq 0.5$)의 산출 빈도에서도 일관되게 확인되었다.

효과 크기 0.2에서 대립가설을 지지하는 데 필요한 표본크기

그림 1을 살펴보면 효과 크기가 작은 경우(d = 0.2) 모든 조건에서 두 방법의 베이즈 인자가 모두 -0.5에서 0사이로 참고 수준에서 귀무가설을 선호하였다. 이는 작은 효과 크기를 탐지하기에 지금까지 검토한 표본크기(최대 200)가 충분하지 않았음을 시사한다.

JZS와 BFGC 방법이 작은 효과 크기(d = 0.2)를 탐지하는 데 필요한 표본크기를 탐색하기 위해, 빈도주의의 검정력 분석 기준을 참조하였다. 베이즈 인자와 p값은 서로 다른 접근과 해석적 틀을 제공하나, 표본크기 계획에서는 검정력 기준이 참고할 만한 출발점을 제공한다(Jeffreys, 1961). 빈도주의에서는 효과 크기 0.2에 대해 유의수준 .05, 검정력 80%를 달성하려면 총 표본크기 600이 필요하다. 유의수준 .01에서 같은 검정력을 달성하려면 표본크기 1000이 필요하다(Cohen, 1988). 빈도주의 검정력 분석 기준에 따라 표본크기 600과 1000 조건에서 베이즈 인자가 대립가설을 선호하는지 확인하기 위한 시뮬레이션을 추가로 수행하였다.

총 표본크기 600과 1000 조건에서 JZS와 BFGC 방법 모두 양의 베이즈 인자를 산출하여 대립가설을 선호하였다. 총 표본크기가 600일 때는 모든 조건에서 로그 베이즈 인자가 0.5 미만으로 나타나 실질적 증거 수준에는 미치지 못하였다. 총 표본크기가 1000일 때 로그 베이즈 인자의 평균은 모든 조건에서 0.5를 넘었다(그림 2).

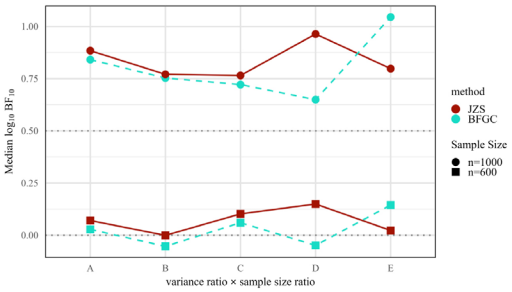

그림 2. 표본크기가 클 때 JZS와 BFGC 사전분포 설정의 조건별 로그 베이즈 인자의 평균(d = 0.2)

논 의

2000년대 이후 심리학 연구에서는 베이지안 방법론의 활용이 꾸준히 증가하고 있다(van de Schoot et al., 2017; Heck et al., 2022). *Psychological Science*는 필요에 따라 베이지안 접근법과 베이즈 인자를 적극적으로 사용할 것을 권장하고 있으며(Association for Psychological Science, 2024) 베이지안 접근법은 빈도주의 접근과 더불어 점차 심리학 연구의 주요 통계 도구로 자리매김하고 있다.

그러나 베이지안 방법론의 확산에도 불구하고, JZS와 같이 자주 권장되는 사전분포의 설정이 잘못되었을 때 가설검정의 결과에 어떠한 영향을 미치는지는 알려진 것이 많지 않았다. 본 연구는 베이지안 t 검정에서 기본 사전분포로 권장되는 JZS 사전분포 기반 모형의

성능을 다양한 조건에서 평가하고 JZS와 BFGC 방법을 비교하여 분산 가정에 따른 차이점을 확인할 수 있었다.

JZS와 BFGC 사전분포 기반 베이즈 인자는 분산과 표본크기의 조합에 따라 상반된 패턴을 보였는데, 이러한 차이는 각 검정의 분산 가정 구조에서 기인한다. 이는 통계학의 베렌스-피셔 문제와 관련된 공통된 속성이다. 등분산을 가정하는 JZS 사전분포는 두 집단의 관측치를 분산 모수 σ^2을 공유하는 정규분포로 정의하며, σ^2에 제프리스 사전분포를 부여하여 주변화한다. 이러한 구조는 결과적으로 전통적인 t 통계량과 합동분산(pooled variance), 그리고 유효 표본크기 $\nu = n_1 n_2/(n_1 + n_2)$를 사용하여 베이즈 인자를 계산한다(Rouder et al., 2009).

합동분산은 각 집단의 표본크기에 비례하여 각 분산을 가중 평균한다. 이분산 조건에서 분산이 큰 집단의 표본크기가 상대적으로 크면 등분산 가정하에 추정되는 합동분산의 크기는 큰 분산 모수 쪽으로 편향된다. 따라서 이 조건에서는 JZS 기반 베이즈 인자 계산에 사용되는 합동분산이 과대 추정되며, 표준화 효과 크기(d =μ/σ)는 과소 추정되고, 결과적으로 대립가설에 대한 증거는 약화된다. 반대로 분산이 큰 집단의 표본크기가 상대적으로 작으면 합동분산의 크기는 작은 분산 쪽으로 편향되어 과소 추정된다. 결국 표준화 효과 크기는 과대 추정되고 베이즈 인자는 대립가설을 상대적으로 강하게 지지한다. 이는 Student의 t 검정에서 등분산 가정 위반 시 나타나는 편향과 동일한 메커니즘이다.

한편, BFGC 사전분포는 효과 크기(Δ)의 사후분포 분산이 $s_1^2/n_1 + s_2^2/n_2$로 독립적인 두 사전분포로부터 계산된다. 이분산 조건에서 분산이 큰 집단의 표본크기가 상대적으로 크면 우도의 영향이 커져 해당 집단 분산의 사후 불확실성이 감소한다. 결과적으로 효과 크기의 전체 불확실성이 감소하여 대립가설에 대한 증거가 강화된다. 반면, 분산이 큰 집단의 표본크기가 상대적으로 작다면, 불확실성이 증가하여 대립가설에 대한 증거가 약화된다. 이는 Welch의 t 검정과 동일한 속성을 공유한다. 실제로 BFGC 사전분포의 사후분포 분산은 Welch의 t 검정과 동일한 형태를 가진다. 이는 BFGC의 사후분포가 Welch의 t 검정이 근사하는 베렌스-피셔 분포에 기반한다는 점에서도 이론적으로 일관된다.

본 연구의 결과는 연구자가 베이지안 t 검정을 적용할 때 분산과 표본크기의 특성을 고려하여 적절한 모형을 선택해야 함을 시사한다. JZS 사전분포를 사용할 경우, 특히 이분산 조건에서 분산이 큰 집단의 표본크기가 클 때 베이즈 인자가 대립가설을 상대적으로 약하게 지지할 수 있음을 유의해야 한다. 중요한 점은 이분산 조건에서 JZS 사전분포를 사용하는 베이지안 t 검정이 부정확한 결과를 제공하는 것이 아니라, 분산과 표본크기 비율의 상호작용에 따라 특정 가설을 체계적으로 선호하는 경향을 보인다는 것이다. 이러한 현상은 모형의 등분산 가정과 실제 데이터 구조 간의 불일치에서 기인한 체계적 패턴으로 이해하는 것이 타당하다.

본 결과는 베이지안 t 검정 적용 시 다음과 같은 실천적 지침을 제시한다. 두 집단 간 분산 차이나 표본크기 불균형이 존재하는 상황에서는 JZS 기반 베이지안 t 검정의 결과를 신

중하게 해석할 필요가 있다. 특히 분산이 큰 집단의 표본크기가 상대적으로 크거나 작은 경우, JZS 모형이 체계적으로 특정 가설을 선호할 수 있음을 고려해야 한다. 따라서 이분산·표본크기 불균형 조건에서는 이분산을 허용하는 베이지안 모형(예: BFGC 사전분포)과의 비교를 통해 결과의 강건성을 확인하기를 권장한다. 이러한 지침은 베이지안 t 검정의 적절한 적용에 유용한 근거를 제공하나, 몇 가지 한계점을 함께 고려할 필요가 있다.

먼저 본 연구는 이분산 조건에서 표본크기 불균등이 오류율을 증가시킴을 확인하였으나(대비 3), 이 효과는 등분산 조건에서는 일부 경우에만 나타났으며(대비 2: d = 0.5, N = 200) 일관성이 낮았다. 이는 표본크기 불균등이 독립적 요인이라기보다 분산·표본크기 교차 효과의 일부로 작동함을 시사한다. 미세한 차이(1~4%p)가 시뮬레이션 오차인지 체계적 패턴인지 판단하기 위해서는 추가적인 베이지안 t 검정 모형을 포함한 비교를 통해 확인이 필요하다.

나아가 베이지안 t 검정에는 본 연구에서 다룬 JZS와 BFGC 사전분포 외에도 부분 베이즈 인자(fractional Bayes factor; Moreno et al., 1999), MCMC 기반 접근법(Wetzels et al., 2009), 모형 평균화 베이지안 t 검정(Maier et al., 2024) 등 다양한 방법론이 존재한다. 본 연구 결과는 JZS와 BFGC가 분산 가정에 따라 상이한 패턴을 보임을 확인하였으며, 이는 다른 베이지안 접근법들도 유사한 체계적 차이를 보일 가능성을 시사한다. 따라서 후속 연구에서는 이들 방법론을 포괄적으로 비교하여 등분산 가정에 따른 베이즈 인자 산출 패턴의 공통점과 차이점을 탐색할 필요가 있다. 이를 통해 연구자들이 자신의 연구 맥락에 가장 적합한 베이지안 방법론을 선택할 수 있는 경험적 근거를 제공할 수 있을 것이다.

참고문헌

김달호. (2013). R과 WinBUGS를 이용한 베이지안 통계학 (2판) [Bayesian statistics (Using R and WinBUGS) (2nd ed.)]. 자유아카데미.

Association for Psychological Science. (2024, December 6). Psychological Science: Submission guidelines. Retrieved April 8, 2025, from https://www.psychologicalscience.org/publications/psychological_science/ps-submissions

Behrens, W. U. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen [A contribution to error estimation with few observations]. *Landwirtschaftliche Jahrbücher*, *68*, 807-837.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*(3), 317-335. https://doi.org/10.1214/ss/1177013238

Berger, J. O., & Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, *91*(433), 109-122. https://doi.org/10.1080/01621459.1996.10476668

Borchers H (2023). _pracma: Practical Numerical Math Functions_. R package version 2.4.4. https://CRAN.R-project.org/package=pracma

Box, G. E., & Tiao, G. C. (1973). Bayesian inference in statistical analysis. Addison-Wesley.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental

studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*(3), 396-404. https://doi.org/10.1037/0033-2909.104.3.396

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. L. Erlbaum Associates.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997-1003. https://doi.org/10.1037/0003-066X.49.12.997

Davison, A. C. (2001). Biometrika Centenary: Theory and general methodology. *Biometrika*, *88*(1), 13-52. https://doi.org/10.1093/biomet/88.1.13

Dayal, H. H., & Dickey, J. M. (1976). Bayes factors for Behrens-Fisher problems. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, *38*(4), 315-328. http://www.jstor.org/stable/25052031

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, *30*(1), 92-101. https://doi.org/10.5334/irsp.82

Fisher, R. A. (1935). The fiducial argument In statistical inference. *Annals of Eugenics*, *6*(4), 391-398. https://doi.org/10.1111/j.1469-1809.1935.tb02120.x

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/b16018

Girón, F. J., & Del Castillo, C. (2021). A Bayesian solution to the Behrens-Fisher problem. *Revista De La Real Academia De Ciencias Exactas Físicas Y Naturales Serie a Matemáticas*, *115*(4). https://doi.org/10.1007/s13398-021-01098-0

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-Sample t test. *The American Statistician*, *59*(3), 252-257. https://doi.org/10.1198/000313005X55233

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2018). Comparing Objective and Subjective Bayes Factors for the Two-Sample Comparison: The Classification Theorem in Action. *The American Statistician*, *73*(1), 22-31. https://doi.org/10.1080/00031305.2017.1322142

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337-350. https://doi.org/10.1007/s10654-016-0149-3

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2019). Informed Bayesian t-tests. *The American Statistician*, *74*(2), 137-143. https://doi.org/10.1080/00031305.2018.1562983

Gu, X., Mulder, J., & Hoijtink, H. (2017). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229-261. https://doi.org/10.1111/bmsp.12110

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. a. L., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A.,

Meijerink-Bosman, M., Moerbeek, M., . . . Hoijtink, H. (2022). A review of applications of the Bayes factor in psychological research. *Psychological Methods*, *28*(3), 558-579. https://doi.org/10.1037/met0000454

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539-556. https://doi.org/10.1037/met0000201

IBM Corp. (2023). IBM SPSS Statistics for Windows (Version 29.0) [Computer software].

JASP Team (2025). *JASP* (Version 0.95.4) [Computer software].

Jeffreys, H. (1937). On the relation between direct and inverse methods in statistics. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *160*(902), 325-348. http://www.jstor.org/stable/96806

Jeffreys, H. (1940). Note on the Behrens Fisher formular. *Annals of Eugenics*, *10*(1), 48-51. https://doi.org/10.1111/j.1469-1809.1940.tb02236.x

Jeffreys, H. (1961). Theory of probability (3rd ed.). Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795. https://doi.org/10.1080/01621459.1995.10476572

Kelter, R. (2020). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics*, *36*(2), 1263-1288. https://doi.org/10.1007/s00180-020-01034-7

Kelter, R. (2021). A New Bayesian Two-Sample t Test and Solution to the Behrens-Fisher Problem Based on Gaussian Mixture Modelling with Known Allocations. *Statistics in Biosciences*, *14*(3), 380-412. https://doi.org/10.1007/s12561-021-09326-2

Kim, S.-H., & Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics*, *23*(4), 356-377. https://doi.org/10.2307/1165281

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573-603. https://doi.org/10.1037/a0029146

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259-269. https://doi.org/10.1177/2515245918770963

Lenth, R. V. (2025). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.11.0. https://CRAN.R-project.org/package=emmeans

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19-32. https://doi.org/10.1016/j.jmp.2015.06.004

Maier, M., Bartoš, F., Quintana, D. S., Dablander, F., Van den Bergh, D., Marsman, M., Ly, A., & Wagenmakers, E.-J. (2024). Model-averaged Bayesian t tests. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-024-02590-5

Moreno, E., Bertolino, F., & Racugno, W. (1999b). Default Bayesian analysis of the Behrens-Fisher

problem. *Journal of Statistical Planning and Inference*, *81*(2), 323-333. https://doi.org/10.1016/s0378-3758(99)00070-1

Morey R, Rouder J (2024). BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12-4.7. https://CRAN.R-project.org/package=BayesFactor

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225-237. https://doi.org/10.3758/pbr.16.2.225

Satterthwaite, F. E. (1946), An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110-114. https://doi.org/10.2307/3002019

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2023). What is a Bayes factor? *Psychological Methods*, *28*(3), 705-718. https://doi.org/10.1037/met0000421

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2024). afex: Analysis of Factorial Experiments. R package version 1.4-1. https://CRAN.R-project.org/package=afex

Stanton, J. M. (2017). *Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R*. Guilford Publications.

Student. (1908). The Probable error of a mean. *Biometrika*, *6*(1), 1-25. https://doi.org/10.2307/2331554

The jamovi project (2025). *jamovi* (Version 2.6) [Computer Software].

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217-239. https://doi.org/10.1037/met0000100

Van Ravenzwaaij, D., & Etz, A. (2021). Simulation studies as a tool to understand Bayes factors. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920972624

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158-189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169-176. https://doi.org/10.1177/0963721416643289

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Rouder, J. F., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35-57. https://doi.org/10.3758/s13423-017-1343-3

Welch, B. L. (1938). The significance of the difference between two means when the

population variances are unequal. *Biometrika*, *29*(3/4), 350-362.
https://doi.org/10.2307/2332010

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, *34*(1/2), 28-35.
https://doi.org/10.2307/2332510

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*(4), 752-760.
https://doi.org/10.3758/pbr.16.4.752

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology. *Perspectives on Psychological Science*, *6*(3), 291-298.
https://doi.org/10.1177/1745691611406923

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos De Estadistica Y De Investigacion Operativa*, *31*(1), 585-603.
https://doi.org/10.1007/bf02888369

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 173-181.
https://doi.org/10.1348/000711004849222

1차원고접수 : 2025. 12. 16
2차원고접수 : 2026. 01. 28
최종게재결정 : 2026. 02. 02

한국심리학회지: 일반
Korean Journal of Psychology: General
2026, Vol. 45, No. 1, 31-54
http://dx.doi.org/10.22257/kjp.2026.3.45.1.31

Bayes Factors for Independent Samples t-Test Under Equal and Unequal Variance Assumptions

Chaereen Kang **Seungmin Jahng**

Department of Psychology, Sungkyunkwan University

Heterogeneity of variance is a persistent concern in independent-samples *t* tests, raising questions about the robustness of Bayesian hypothesis testing when the equal-variance assumption is violated. The Jeffreys-Zellner-Siow (JZS) prior, commonly used as the default in Bayesian *t* tests, inherently assumes homoscedasticity. The present study examines the implications of this assumption by comparing a homoscedastic Bayesian *t* test based on the JZS prior with a heteroscedastic alternative that allows group-specific variances, the Girón-del Castillo (BFGC) model. An extensive simulation study was conducted to investigate how Bayes factors behave across varying combinations of variance ratios, sample size ratios, standardized effect sizes, and total sample sizes. Particular attention was given to conditions in which sample size imbalance interacted with variance heterogeneity. The results showed that the two models exhibit qualitatively different patterns of evidence accumulation under heteroscedasticity. Specifically, the JZS-based Bayes factor tended to provide weaker support for the true hypothesis when the group with the larger variance also had the larger sample size, whereas the BFGC-based Bayes factor showed the opposite pattern, yielding weaker support when the larger-variance group had the smaller sample size. These findings highlight that variance assumptions in Bayesian *t* tests can systematically influence the interpretation of Bayes factors, especially in the presence of sample size imbalance. When heteroscedasticity is plausible, adopting a heteroscedastic Bayesian model such as BFGC may therefore lead to more reliable Bayesian inference than reliance on the default JZS specification.

Key words : Bayes factor, Bayesian t-test, heteroscedasticity, unequal sample sizes

보충 자료

본문의 평균 막대그래프 외에 시뮬레이션을 통해 산출된 베이즈 인자의 추가분석 자료를 보충 자료에 제시하였다. 베이즈 인자의 산출 분포를 각 방법과 조건 별로 상자그림으로 표현하였으며, 시뮬레이션으로 생성된 데이터의 표본평균 차이를 중심으로 두 베이즈 인자의 차이를 살펴보았다. 모든 베이즈 인자는 상용로그로 변환하여 분석되었다.

1. 제1·2종 오류율 산출 패턴 그림

그림 S.1은 본문 결과 표 4와 5에 제시된 오류율 그림을 본문 그림 1과 같은 방식으로 표현한 것으로, 각 표본크기 패널에 분산비 및 표본크기의 조합과 효과 크기, 방법에 따른 오류율을 나타낸다. 귀무가설이 참인 조건(빈 원)에서는 매우 낮은 제1종 오류율(<0.02)을 보였다. 대립가설이 참인 조건(d = 0.2, 0.5, 0.8)에서는 그림 1의 평균 베이즈 인자와 반대 방향의 패턴이 나타났다. 즉, 제2종 오류율은 효과 크기가 작을수록 높았으며, 효과 크기와 표본크기가 증가할수록 감소하였다. 주목할 점은 효과 크기 0.5 이상에서 조합 D와 E 간 교차 패턴이 평균 베이즈 인자와 일치하게 나타났다는 것이다. JZS는 조합 D에서, BFGC는 조합 E에서 더 낮은 제2종 오류율을 보였다(본문 결과 참조).

2. Box plot

그림 S.2부터 S.4는 각각 두 집단 표본크기의 총합(N)이 50, 100, 200일 때 시뮬레이션을 통해 생성된 베이즈 인자의 분포이다. 각 패널은 효과 크기 조건 0, .2, .5, .8 별로 표기하였다. 조건 1은 두 집단의 표본크기가 같고(1:1) 등분산인 경우, 조건 2는 두 집단의 표본크기가 같고(1:1) 이분산(2:1)인 경우, 조건 3은 두 집단의 표본크기가 다르고(2:3) 등분산인 경우, 조건 4는 두 집단의 표본크기가 다르고(2:3) 이분산인 경우(2:1), 마지막으로 조건 5는 두 집단의 표본크기가 다르고(3:2) 이분산인 경우(2:1)이다. 이는 분산비와 표본크기비의 조합 A ~ E와 같다.

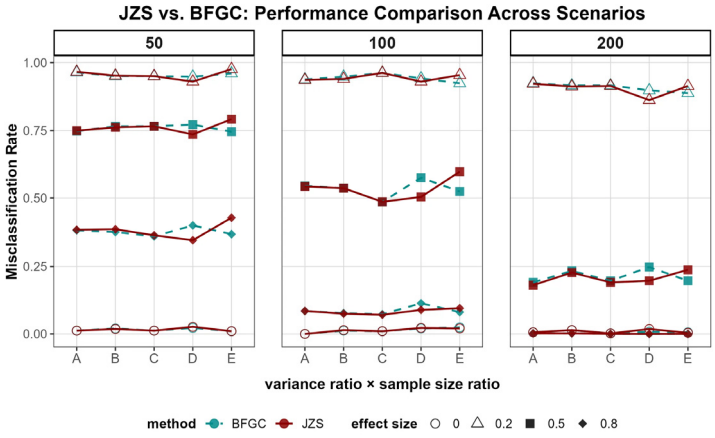

그림 S.1. $\log_{10}BF_{10}$ = 0.5를 기준으로 산출한 베이즈 인자 가설검정 오류율

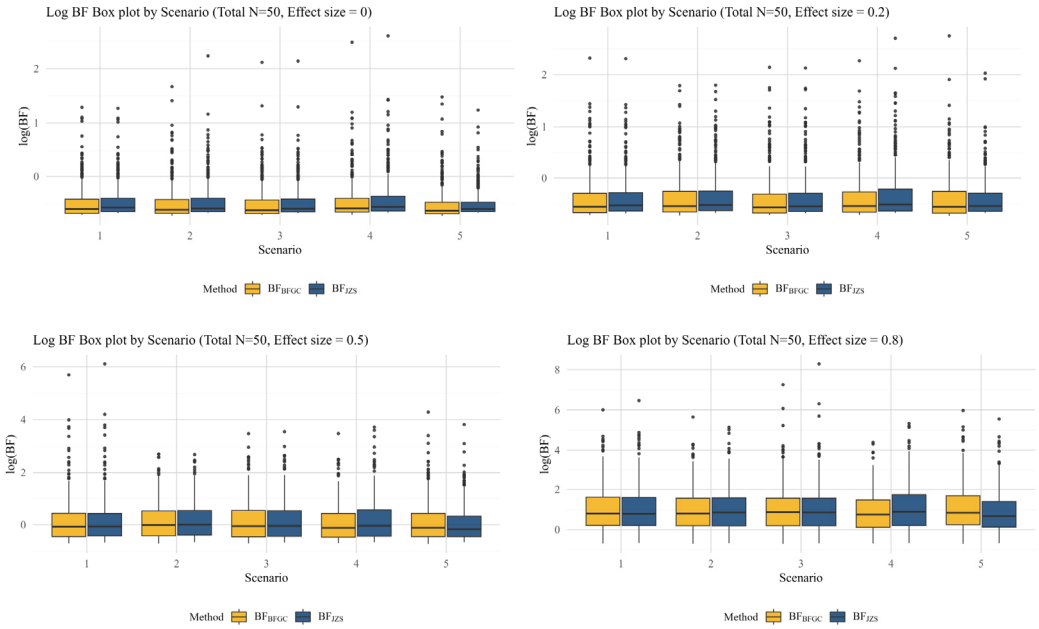

그림 S.2. 효과크기에 따른 시나리오별 로그 베이즈 인자 분포(N = 50)

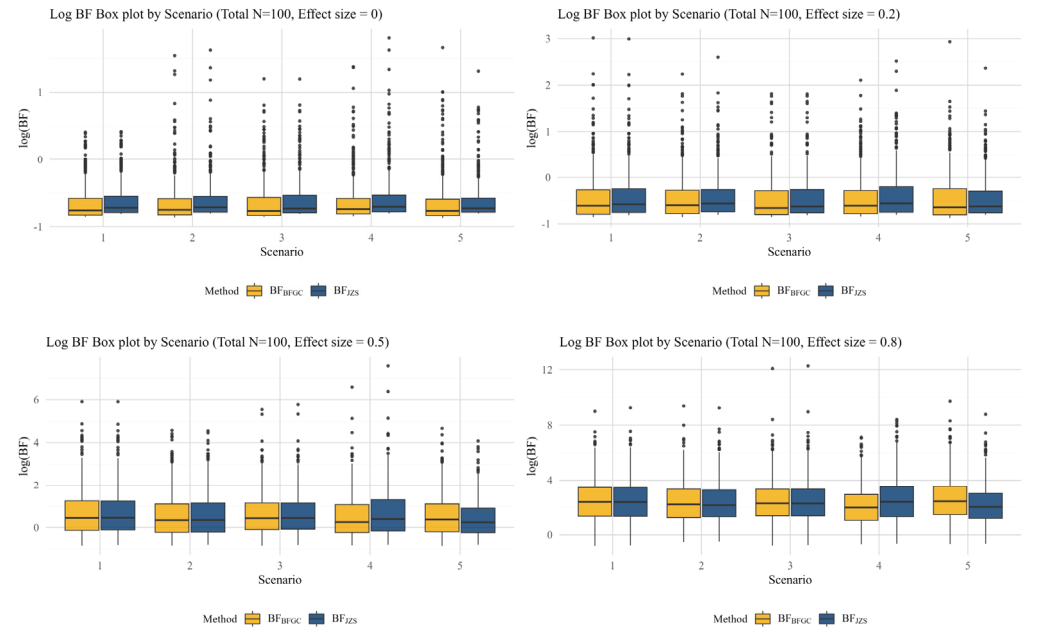

그림 S.3. 효과크기에 따른 시나리오별 로그 베이즈 인자 분포(N = 100)

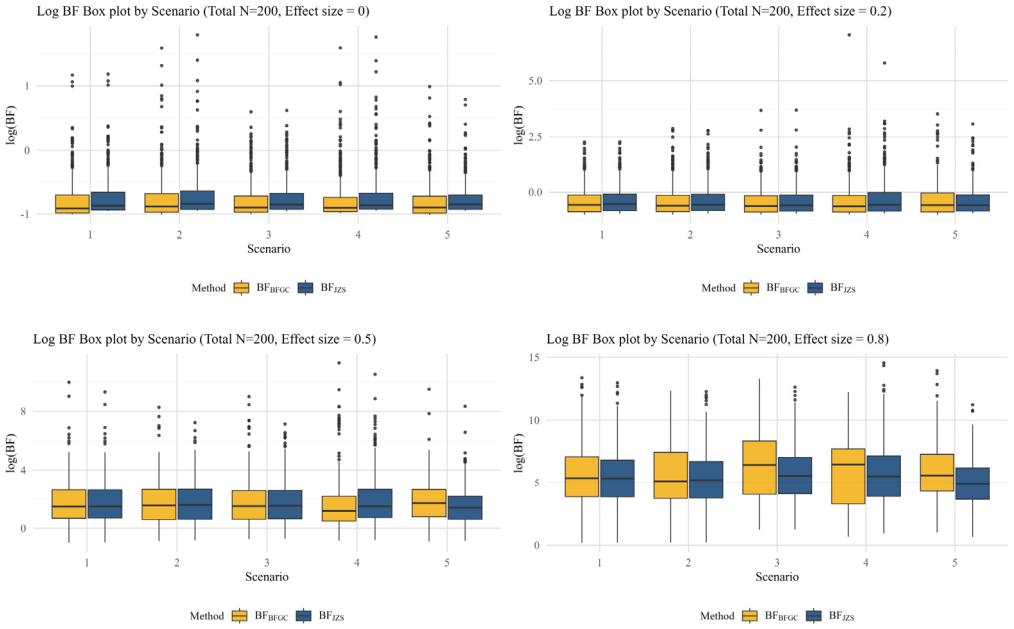

그림 S.4. 효과크기에 따른 시나리오별 로그 베이즈 인자 분포(N = 200)