p-value

> **The ASA's Statement on *p*-Values: Context, Process, and Purpose - Ronald L. Wasserstein**

> GPT

한줄요약

- p-value: 귀무가설이 맞다고 가정했을 때 이런 데이터가 나올 확률

정의

- 확률 값: p-value는 0과 1 사이의 값을 가지는 확률입니다.
- 쉽게 말해, p-value는 특정 통계 모델 하에서 데이터의 통계적 요약 (예: 두 비교 그룹 간의 표본 평균 차이)이 관찰된 값과 같거나 더 극단적일 확률입니다.
    - 특정 통계 모델 하에서 ⇒ 일반적으로 이 "특정 통계 모델"은 **귀무 가설(Null Hypothesis)**이 참이라는 가정(예: '효과가 없다', '그룹 간 차이가 없다')을 의미합니다. 즉, 우리가 보고자 하는 현상이나 효과가 실제로는 없다고 가정했을 때를 전제로 합니다.
    - 데이터의 통계적 요약 ⇒ 이는 연구에서 얻은 데이터에서 계산된 특정 값(예: 두 그룹 평균의 차이, 상관계수 등)을 말합니다.

    - 관찰된 값과 같거나 더 극단적인 값 ⇒ 우리가 실제 실험이나 관찰을 통해 얻은 결과(관찰된 값)만큼이나, 혹은 그보다 더 '놀랍거나' '예상 밖의' 결과가 나올 확률을 의미합니다. 여기서 '극단적'이라는 것은 귀무 가설이 참이라면 관찰되기 어려운 방향을 의미합니다.
- 결론: 따라서 p-value가 작으면 작을수록, 귀무 가설이 참이라는 가정 하에 현재 관찰된 데이터나 그보다 더 극단적인 데이터가 나올 확률이 매우 낮다는 것을 의미합니다. 이는 귀무 가설이 데이터와 잘 맞지 않는다는 (즉, 귀무 가설에 대한 증거가 약하다는) 통계적 불일치를 나타내는 지표가 됩니다.

💡 ✅ 핵심 정리 (예시)

🔷 1. p-value = 1 이라면?

- 귀무가설이 “두 집단은 같다”일 때, 데이터가 그 가설에 **완벽하게 부합**
- 즉, **차이 없다**는 가정이 너무 잘 맞음
- 데이터가 “너무 평범해서” **실험적 차이가 없다는 강력한 신호**

👉 하지만 **“100% 같다”는 의미는 아님**.
실제로는 **“차이가 있다는 증거가 없다”**라는 말이 더 정확해요.

🔷 2. p-value = 0 이라면?

- 귀무가설이 참일 때 이런 데이터는 **절대 나올 수 없을 정도로 극단적**
- 즉, 귀무가설(두 집단이 같다)을 **강하게 부정**
- 데이터는 **거의 확실히 다르다**

👉 이건 “100% 다르다”는 의미도 아니고,
“우연으로 이렇게 다를 확률이 거의 0”이라는 뜻이에요.

원칙 1: p-값은 데이터가 특정 통계 모델과 얼마나 양립 불가능한지를 나타낼 수 있습니다.

p-값은 주어진 데이터 세트가 제안된 모델(종종 '효과가 없다'는 귀무 가설)과 얼마나 잘 맞지 않는

원칙 2: p-값은 연구된 가설이 참일 확률이나 데이터가 순전히 우연에 의해 생성되었을 확률을 측정하지 않습니다.

연구자들은 종종 p-값을 귀무 가설의 진실성이나 데이터가 우연히 발생했을 확률로 오해합니다. 하지만 p-값은 특정 가상의

원칙 3: 과학적 결론 및 비즈니스 또는 정책 결정은 p-값이 특정 임계값을 통과했는지 여부에만 근거해서는 안 됩니다.

"p < 0.05"와 같은 기계적인 '명확한 기준' 규칙에 따라 데이터 분석이나 과학적 추론을 단

지를 요약하는 한 가지 방법입니다.
p-값이 작을수록 데이터가 귀무 가설과 통계적으로 더 많이 불일치함을 의미하며, 이는 귀무 가설이나 p-값 계산에 사용된 기본 가정에 대한 의문을 제기하는 증거가 될 수 있습니다.

원칙 4: 적절한 추론을 위해서는 완전한 보고와 투명성이 필요합니다.

p-값과 관련 분석은 선택적으로 보고되어서는 안 됩니다. 특정 p-값을 얻기 위해 여러 분석을 수행하고 일부 결과만(일반적으로 유의성 임계값을 통과한 결과만) 보고하는 것은 보고된 p-값을 해석 불가능하게 만듭니다."데이터 드레징", "p-해킹" 등으로 알려진 유망한 결과만 선별적으로 고르는 행위는 학술 문헌에 통계적으로 유의미한 결과가 과도하게 넘쳐나게 하므로 강력히 피해야 합니다. 연구자는 연구 중에 탐색한 가설의 수, 모든 데이터 수집 결정, 수행한 모든 통계 분석, 설명(귀무 가설)과 관련하여 데이터가 얼마나 극단적인지에 대한 진술일 뿐, 그 설명 자체에 대한 진술이 아닙니다.

원칙 5: p-값 또는 통계적 유의성은 효과의 크기나 결과의 중요성을 측정하지 않습니다.

통계적 유의성은 과학적, 인간적 또는 경제적 중요성과 동일하지 않습니다. p-값이 작다고 해서 반드시 더 크거나 더 중요한 효과가 있음을 의미하지 않으며, p-값이 크다고 해서 중요성이 없거나 효과가 없음을 의미하지도 않습니다. 표본 크기가 충분히 크거나 측정 정밀도가 높으면 아무리 작은 효과라도 작은 p-값을 생성할 수 있으며, 표본 크기가 작거나 측정이 부정확하면 큰 효과도 인상적이지 않은 p-값을 생성할 수 있습니다.

순화하는 관행은 잘못된 믿음과 좋지 않은 의사 결정을 초래할 수 있습니다.
결론은 단순히 임계값 한쪽에서는 "참"이 되고 다른 쪽에서는 "거짓"이 되지 않습니다. 연구 설계, 측정의 질, 연구 중인 현상에 대한 외부 증거, 데이터 분석의 기본 가정 타당성 등 여러 맥락적 요소를 종합적으로 고려해야 합니다.

원칙 6: p-값만으로는 모델이나 가설에 대한 증거의 좋은 척도가 되지 않습니다.

연구자들은 맥락이나 다른 증거 없이 p-값만으로는 제한된 정보만을 제공한다는 점을 인식해야 합니다. 예를 들어, 0.05에 가까운 p-값은 귀무 가설에 대한 약한 증거만을 제공할 뿐입니다.
마찬가지로, 상대적으로 큰 p-값은 귀무 가설에 유리한 증거를 의미하지 않습니다. 다른 많은 가설들이 관찰된 데이터와 동등하거나 더 일관될 수 있습니다. 이러한 이유로, 다른 접근 방식이 적절하고 실현 가능하다면 데이터 분석이 p-값 계산으로 끝나서는 안 됩니다.

계산된 모든 p-값을
공개해야 합니다.

t-statistic: 평균 차이의 상대 크기

s.e.(Standard Error) 표준오차

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{\text{s. e.}(\hat{\beta})}$$

✅ t-statistic 수식 (독립표본 t-검정)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

각 기호 의미:

기호	설명
$\bar{X}_1, \bar{X}_2$	두 집단의 평균
s_1^2, s_2^2	두 집단의 분산 (또는 표준편차 제곱)
n_1, n_2	각 집단의 표본 크기
분모 전체	**표준 오차**(SE, Standard Error)

p-value와 차이 ⇒ p-value는 “그 정도 벗어남이 얼마나 드문가?”

t-statistic은 “얼마나 멀리 벗어났는가?”

p-value 대체 접근 방식

1.추정 중심 방법 (Estimation over Testing): 가설 검정보다는 효과의 크기를 추정하는 데 중점을 둡니다.

1. **신뢰 구간(Confidence Intervals):** 모수(예: 평균, 비율 차이)가 포함될 것으로 예상되는 값의 범위를 제공하여 효과의 크기와 그 불확실성을 보여줍니다.
2. **신용 구간(Credibility Intervals): 베이지안 통계**에서

2.베이지안 방법 (Bayesian Methods): 사전 지식(prior knowledge)과 현재 데이터를 결합하여 가설의 확률을 업데이트하고, 가설의 참 거짓에 대한 직접적인 확률을 제공합니다.

- **베이즈정리**

$$P(\theta \mid x) = \frac{P(x \mid \theta)\, P(\theta)}{P(x)}$$

- **베이지안 방법**

사용되며, 모수가 특정 범위 내에 있을 **사후 확률**을 나타냅니다.

3. **예측 구간(Prediction Intervals):** 새로운 관측값이 나타날 범위에 대한 추정치를 제공합니다.

3.대안적인 증거 측정 (Alternative Measures of Evidence):

1. **우도비(Likelihood Ratios)**: 두 가지 다른 가설 하에서 관찰된 데이터가 얼마나 그럴듯한지(likelihood)를 비교하여 한 가설이 다른 가설에 비해 데이터를 얼마나 더 잘 설명하는지 나타냅니다.
2. **베이즈 팩터(Bayes Factors)**: 베이지안 통계에서 두 가설(모델) 간의 상대적인 증거를 측정하며, 어떤 가설이 데이터를 더 잘 지지하는지 정량화합니다.

$$BF = \frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_0)}$$

BF>1 : H1 확률이 높다
BF<1: H0 확률이 높다

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

4.기타 접근 방식 (Other Approaches):

1. **의사 결정 이론적 모델링(Decision-Theoretic Modeling)**: 통계적 추론을 넘어 결정의 결과(비용, 편익, 손실 등)를 명시적으로 고려하여 최적의 의사결정을 내리는 데 중점을 둡니다.
2. **거짓 발견율(False Discovery Rates, FDR):** 다중 가설 검정 상황에서 잘못된 양성(false positive)으로 분류된 가설의 비율을 통제하는 방법입니다.

우도(likelihood)=가능도

$$L(\theta \mid x) = P(x \mid \theta)$$

- 큰 값일 수록 좋음 (확률을 넣었을때 나오는 값 -**확률이 아님**-)
- 나온데이터가 어떤 확률일때 이런 데이터가 나오기 제일 유력한지.

- 즉. 한 데이터 표본에 대해 여러개 가설이 나온다.
  ex) p=0.1 ⇒likelihood , p=0.2 ⇒liklihood, ...

☑️ ~~p-value는 귀무가설이 맞다고 가정했을 때 이런 데이터가 나올 확률을 의미. likelihood는 어떤 가설이 이 데이터를 잘 설명 하는지.~~

MLE(MAX LIKILIHOOD ESTIMATION) 최대 우도 추정

$$\hat{\theta} = \arg\max_{\theta} L(\theta \mid x)$$

결론

⇒내 주장의 반대가설이라고해서 '반대가설이 틀렸으므로, 내 가설이 옳다'는 잘못된 논리이다.

⇒p-value의 귀무가설은 보통 내 주장과 반대 가설을 두고 p-value를 낮추는 방식으로 많이한다. 보통 p-value<0.05를 많이 사용한다(원칙3:그렇다고 0.05에 의존하는것은 잘못된 것).

⇒p-value 맹신 하지말고 여러가지 수치보면서 판단해라.