# P-value
# 이해하기

# P-value 개념 정의

## 검정의 원리

가설검정의 원리

예) 어느 전구의 제조공정에서는 평균 수명이 $\mu_0 = 1500$(시간)이고, 표준편차가 $\sigma_0 = 100$(시간)이 되도록 품질관리를 하고 있다. 품질을 개선하기 위하여 어느 팀에서는 새로운 공법을 개발하였으며, 새 공법에 의하면 평균 수명이 증가한다고 주장한다. (다만, 새 공법의 특성은 기존의 공법과 같아서 $\sigma = 100$(시간)으로 가정할 수 있으며, 생산비용도 같다고 가정한다.)

이를 확인하기 위하여 $n = 25$개의 전구를 새로운 공법으로 시험 생산하여 시험한 결과 $\overline{x} = 1550$(시간)을 얻었다. 이들 자료로부터 새 공법에 의한 평균 수명이 기존의 공법에 의한 평균 수명보다 길어졌다고 확신할 수 있는가?

① 가설의 설정

$\mu$ : 새로운 공법에서의 전구의 평균수명
$\mu_0(=1500)$ : 기존의 공법에서의 전구의 평균수명

귀무가설과 대립가설 :
$H_0 : \mu = 1500$ (두 공법의 평균 수명은 같다, **“귀무가설”**)
$H_1 : \mu > 1500$ (새 공법의 평균 수명이 기존 공법의 평균 수명보다 길어졌다, **“대립가설”**)

② 자료로부터 $H_0$와 $H_1$ 가운데 하나를 채택하는 과정 (기각역을 이용하는 방법)

$X_1, X_2, ..X_n$ : 랜덤추출한 $n$개의 전구들의 수명시간 (모집단에서의 랜덤표본)
$x_1, x_2, ..., x_n$ : $X_1, X_2, ..X_n$의 관측값 (※ $n=25$, $\overline{x}=1550$)

귀무가설 $H_0$하에서 (즉, $H_0$가 참일 때) $\overline{X}$의 표준화된 통계량 $Z$와 그 관측값 $z$:

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{\overline{X}-1500}{100/\sqrt{25}} \sim N(0,1)$$

$$z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}} = \frac{1550 - 1500}{100 / \sqrt{25}} = 2.5$$

유의수준 5%에서 기각역 : $z \geq z_{0.05} = 1.645$

기각역의 의미 :

통계량 $Z$ 의 값이 $z_{0.05} = 1.645$보다 클 확률은 0.05이다. 즉, $H_0$하에서 $\overline{X}$의 값이 1500보다 훨씬 커져서 통계량 $Z$의 값이 1.645보다 커질 확률이 0.05(5%)이며, 이 경우에는 $H_0$ 보다는 $H_1$이 참이라고 주장할 근거가 있는 셈이다.

결론 : 관측값 $z = 2.5$이므로 기각역에 들어간다. 따라서, $H_0$를 기각하고 $H_1$을 채택하여 새 공법에서 평균수명이 증가했다고 확신하게 된다.

③ 자료로부터 $H_0$와 $H_1$ 가운데 하나를 채택하는 과정 (유의확률을 이용하는 방법)

$X_1, X_2, ..X_n$ : 랜덤추출한 $n$개의 전구들의 수명시간 (모집단에서의 랜덤표본)

$x_1, x_2, ..., x_n$ : $X_1, X_2, ..X_n$의 관측값 (※ $n = 25$, $\overline{x} = 1550$)

귀무가설 $H_0$하에서 (즉, $H_0$가 참일 때) $\overline{X}$의 표준화된 통계량 $Z$와 그 관측값 $z$:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{\overline{X} - 1500}{100 / \sqrt{25}} \sim N(0,1)$$

$$z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}} = \frac{1550 - 1500}{100 / \sqrt{25}} = 2.5$$

귀무가설 $H_0$하에서 (즉, $H_0$가 참일 때) 유의확률 ($p$-value)

$$P(\overline{X} \geq \overline{x}) = P(\overline{X} \geq 1550) = P\left( \frac{\overline{X} - 1500}{100 / \sqrt{25}} \geq \frac{1550 - 1500}{100 / \sqrt{25}} \right) = P(Z \geq 2.5) = 0.0062$$

유의확률 ($p$-value) 값의 의미 :

귀무가설 $H_0$하에서 통계량 $\overline{X}$가 실제 관측된 값 $\overline{x}$보다 대립가설을 지지하는 방향으로 더욱 치우칠 확률이다. 유의확률($p$-value)이 작을수록 $H_0$에 대한 반증이 강한 것을 의미한다.

결론 : 유의수준 5%(0.05)보다 유의확률($p$-value) 값 0.0062이 작다. 따라서, $H_0$를 기각하고 $H_1$을 채택하여 새 공법에서 평균수명이 증가했다고 확신하게 된다.

## 가설검정

가설검정 : 검정통계량을 이용하여 귀무가설 $H_0$의 기각 여부(즉, 대립가설 $H_1$의 채택 여부)를 판정하는 과정

가설검정의 용어 :

대립가설 (alternative hypothesis, $H_1$) : 자료로부터의 강력한 증거에 의하여 입증하고자 하는 가설

귀무가설 (null hypothesis, $H_0$) : 대립가설에 대한 확실한 근거가 없을 때 받아들이는 가설, 기존의 가설

검정통계량 (test statistic) : 가설검정에 사용되는 통계량, 예) $\overline{X}$

유의수준 (significance level) : 미리 정해놓은 기준값

기각역 (rejection region): 귀무가설을 기각시키는 검정통계량의 관측값의 영역

유의확률 : 귀무가설 하에서 검정통계량이 실제 관측된 값보다 대립가설을 지지하는 방향으로 더욱 치우칠 확률

# P-value 이해하기[1]

P-value : 귀무가설 하에서, 관찰된 통계량 만큼의 극단적인 값이 관찰될 확률
주어진 데이터가 얼마나 가능한지의 확률

p-value=0.001 : 귀무가설 하에서 주어진 데이터만큼 극단적인 데이터가 관측되는 것은 1000번에 1번 관측된다는 의미. 즉 거의 불가능하다는 뜻.

The ASA's statement on P-values: context, process, and purpose

1. P-값은 (통계적 유의성보다는) 가정된 모형이 데이터와 별로 맞지 않음을 나타낼 수 있다.
2. P-값은 주어진 가설이 참인 확률이나, 데이터가 랜덤하게 생성된 확률이 아니다.
3. 과학적 연구 결과와 비즈니스, 정책 결정 과정은 P-값이 어떤 경계값보다 크거나 작은 것이 근거해서는 안된다.
4. 제대로 된 추론을 위해서는 연구과정 전반에 대한 보고서와 투명성이 필요하다.
5. P-값이나 통계적 유의성은 효과의 크기나 결과의 중요성을 나타내지 않는다.
6. P-값 자체만으로는 모형이나 가설에 대한 증거가 되지 못한다.

1) 실리콘밸리 데이터 과학자가 알려주는 따라하며 배우는 데이터 과학(권재명) 참고