6. 데이터 기반 의사결정

임채영
서울대학교 통계학과

이번 강의에서 다룰 내용

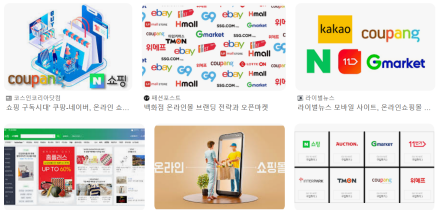
- 유의성 검정의 개념
- 두 그룹의 비교 - A/B 검정

유의성 검정의 개념

유의성 검정 (가설 검정)

- 과학적 의사결정의 방법으로 기존의 이론이나 법칙을 부정하는 것으로 보이는 현상이 관측되었을 때, 이를 유지할지 부정할지를 결정하는데 사용한다.
- 반증을 찾기 위해 설정된 가설 (주로 '기존의 가설'): 귀무가설(Null hypothesis, H_0)
- 귀무가설의 대안으로 상정되는 가설: 대립가설(Alternative hypothesis, H_1)
- 귀무가설에 대한 반증의 강도를 확률을 바탕으로 제공하는 과정을 통해 의사결정을 내리는 것을 유의성 검정(Significance Test)이라 한다.

예제 **1**



- 쇼핑몰 앱 UI를 변경하여 매출을 증가시키고자 할 때 경영자와 웹디자이너가 개인적인 역량으로 의사결정을 할 수 있지만, 데이터에 기반한 의사결정을 할 수 있다.
- 예를 들어, 일부 방문객에 새 디자인을 제공하여 기존 디자인과의 매출 차이를 볼 수 있다. 만약, 7일 동안의 매출액 기준으로 기존보다 일 평균 매출이 100만원 증가하였다면 새 디자인이 매출을 증가시켰다고 할 수 있는가?

- H_0: 기존 디자인과 매출에서 차이가 없다
- H_1: 새 디자인에서 매출이 증가하였다

예제 **2**

- A는 동전을 직접 여러번 던져서 나온 앞면과 뒷면 결과를 순서대로 적게 하는 경우와 머리속으로 가상으로 동전던지기 결과를 적은 경우를 구분 할 수 있다고 주장한다. A는 두 그룹을 구분할 수 있는 능력자 (달인)인가?
- 실제로 달인인지 확인하기 위해 8명의 일반인을 임의로 뽑아 4명에게는 직접 동전을 30번 던져서 나온 결과를 순서대로 적게 하고, 나머지 4명은 가상으로 생각한 동전던지기 결과를 적으라고 하고 그 결과를 A에게 보여주었다.
- A는 결과를 토대로 직접 던진 사람 4명을 맞게 찾아내었다면 A는 달인인가?

- H_0: 달인이 아니다
- H_1: 달인 이다

유의성 검정의 개념

- 예제 2를 가지고 유의성 검정의 개념과 절차를 소개한다.
- 달인의 능력에 대한 기준은 사람마다 다를 수 있고 이에 따라 의사결정이 달라질 수 있다.
- Fisher는 의사결정의 주관성을 제거하기 위해 달인이 아닌 보통사람의 능력을 판단기준으로 생각하는것을 제안 하였다.
- 즉, 보통사람이 가진 능력을 기준으로 매우 일어나기 힘든 결과를 얻었다면 그 결과는 보통사람에 의한 것이 아닌 달인의 능력이라고 보는것이다.

- 보통사람이 가진 능력을 기준으로 실험의 결과가 나타날 확률, 즉 직접 동전을 던진 사람 4명을 모두 맞출 확률이 미리 정한 수준보다 작으면 달인으로 인정하자는 것이다.
- 이때 해당 확률을 p-값(p-value, 유의확률)라고 부르며, 미리 정한 수준을 유의수준(significance level) 이라고 부른다.
- p-값은 보통사람의 기준으로 현재 일어난 사건 또는 더 희박하게 일어날 사건들의 확률로 볼 수 있다.
- 유의수준이란 보통사람의 기준으로 희박하게 일어나는 가능성의 기준을 의미한다.
- 통상적으로 5%, 즉 확률 0.05를 사용하지만 분야마다 기준이 다를 수 있다.

- 보통 사람은 두 가지 경우를 구분할 수 없다고 보는것이 타당하므로, 8명 중에 4명을 임의로 고르는것과 같다.
- 이 경우, 정확하게 4명을 다 골라낼 확률은 $1/\binom{8}{4} = 0.0143$이다.
- 따라서, 유의수준 5%에서 이 사람은 달인으로 인정할 만 하다 (대립가설 선택)는 결론을 내릴 수 있다.

검정통계량

- 의사결정의 결론, 즉 가설의 선택은 p-값을 기준으로 할 수도 있고, p-값을 계산하는데 사용한 실험의 결과값 (이 경우 몇명을 맞게 선택했는지)을 기준으로 할 수도 있다. 이를 검정통계량(Test Statistics)이라고 한다.
- 유의수준 5%에서는 4명을 다 맞추어야, 즉 검정통계량이 4 이상이어야 귀무가설 (보통사람)을 버리고 대립가설 (달인)을 선택하는것이다.
- 귀무가설을 버리는(대립가설 채택, 즉 달인 인정) 검정통계량 값들의 집합(범위)을 기각역(Rejection Region)이라고 한다.

- B는 같은 결과를 보고 직접 동전을 던져서 기록한 4명중 3명을 맞게 골라내고 1명을 잘못 선택하였다고 하자.
- 4명중 3개명을 제대로 선택할 확률은 $\binom{4}{3}\binom{4}{1}/\binom{8}{4} = 16/70 = 0.2286$이며 4명을 모두 맞게 선택하는 확률은 앞에서 구한 0.0143이므로 두 확률을 합하여 p-값이 된다.
- 따라서 p-값은 0.2428(=0.2286+0.0143)이고 유의수준 5%에서 B는 달인이라고 할 수 없다.
- 검정통계량이 4이상이어야 귀무가설을 기각하는 것을 기준으로 삼았을때도 B는 달인이 아니라고 할 수 있다.

p-값

- 앞에서 p-값(또는 유의확률)을 귀무가설이 참 일때, 데이터로부터 계산한 검정통계량의 값(검정통계량 관측값) 또는 그보다 더한 (더 잘 안일어나는) 값의 확률로 설명하였다. p-값에 대한 또 다른 설명들은 다음과 같다.

- 검정통계량 관측값을 가지고 귀무가설이 참인데 귀무가설을 기각하게 하는 가장작은 유의수준.
- 검정통계량의 관측값을 포함하는 기각역중에 귀무가설 하에서의 확률을 최소로 하는 기각역의 확률

유의성 검정의 오류

- 유의수준을 5%가 아니고 1%로 정하였다면 처음 실험에서 A는 달인으로 인정 받지 못한다. 즉, 유의수준에 따라 결과가 달라질 수 있다.
- 결과가 달라진다는 것은 유의성 검정이 완벽하지 않다는 의미이다 (오류의 가능성)
- 보통사람이라는 가정(귀무가설이 맞다는 가정)하에 관측한 결과가 나올 (또는 그보다 더 한 결과가 나올) 확률이 미리 정한 유의수준보다 낮으면 달인이라고 인정(귀무가설 기각)하므로, 미리 정한 유의수준은 귀무가설이 참인데 기각할 오류의 상한을 의미한다.
- 즉, 유의수준 5%인 검정은 귀무가설이 참인데 기각할 오류가 5% 이하인 검정방법이다.

- 유의성 검정에 따른 결론이 틀리는 경우는 두 가지가 있다.
- 1종 오류 (type I error): 귀무가설이 옳은 상황에서 귀무가설을 기각함으로 인해 생기는 오류
- 2종 오류(type II error): 귀무가설이 틀린 상황에서 귀무가설을 기각하지 못함으로 인해 생기는 오류

실제 현상 / 검정결과	H_0 참	H_1 참
H_0 채택	옳은 결정	제 2종 오류
H_1 채택	제 1종 오류	옳은 결정

오류가 적은 검정

- 1종 오류가 일어날 확률 (α) = 유의수준 - 귀무가설이 참인데 기각할 오류의 확률
  - 유의수준 5%인 검정은 $\alpha = 0.05$라는 뜻이고, 귀무가설이 참인데 기각할 오류를 5% 이하로 하겠다는 것이다.
- 2종 오류가 일어날 확률 (β) - 귀무가설이 거짓인데 기각하지 않는 오류의 확률
- 검정력 (power) = $1 - \beta$ - 귀무가설이 거짓일때 귀무가설을 기각할 확률
- 두가지의 오류를 동시에 작게 하기 어렵기 때문에, 보통 1종 오류가 일어날 확률을 정하고 (controlling type I error), 그중에 2종 오류가 일어날 확률이 적은 검정법을 고려한다.
- 주어진 유의수준(1종오류의 확률)하에서 검정력이 가장 큰 (2종 오류가 제일 작은) test를 most powerful test라고 부른다.

유의성 검정의 순서

- 요약하면 유의성 검정은 다음의 순서로 진행한다.

(1) 가설 (귀무가설, 대립가설)을 세운다

(2) 유의수준을 정한다

(3) 실험(데이터 수집)을 계획하고 검정통계량을 정한다

(4) 관측된 결과를 이용하여 p-값을 구하여 유의수준과 비교

또는 유의수준에 따른 귀무가설을 기각할 수 있는 검정통계량의 범위(기각역)에 관측한 검정통계량이 포함되는지를 비교하여 결론을 내린다.

유의성 검정에서 알아두어야 할 사항

- 검정통계량이 다르면 다른 검정법
- 기각역의 형태는 대립가설의 영향을 받음
- *p*-값(유의확률) 또는 기각역을 구하기 위해서는 귀무가설하에서 검정통계량의 분포를 알아야 함
- 어떤 가설을 귀무가설로?
  - 기존에 믿어오던(알려져 있던) 사실
  - 오류의 위험이 더 큰 경우를 1종 오류가 되도록 정함

모평균에 대한 유의성 검정 (모분산을 모르는 정규모집단)

- t검정 (one sample t-test)
- 귀무가설: $H_0 : \mu = \mu_0$
- 검정통계량: $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$: 귀무가설하에서 $T \sim t(n-1)$
- 검정통계량의 관측값: t_0

$T \sim t(k)$일 때, p분위수 $t_p(k)$는 $P(T \leq t_p(k)) = p$를 만족하는 값이다.

대립가설 H_1	유의확률	유의수준 α의 기각역
$\mu > \mu_0$	$P(T > t_0)$	$T > t_{1-\alpha}(n-1)$
$\mu < \mu_0$	$P(T < t_0)$	$T < -t_{1-\alpha}(n-1)$
$\mu \neq \mu_0$	$P(\|T\| > \|t_0\|)$	$\|T\| > t_{1-\alpha/2}(n-1)$

예제

A회사는 세척 공정에서 12인치 웨이퍼 한장 당 평균 1500L의 초순수를 사용하고 있다. 여러 단계의 조정을 통해 1500L보다 적게 사용할 수 있는 공정을 개발하였다고 하자. 이를 확인하기 위해 총 30번 공정을 가동하여 웨이퍼 1장당 사용한 초순수 양을 조사하였다. 그 결과 30개의 데이터의 평균과 표준편차는 다음과 같다. $\bar{x} = 1403.00$, $s = 151.57$
적당한 가설을 세우고, 유의수준 5%, 에서 가설을 검정해보자.

```python
# 요약 통계
n = len(x)
xbar = x.mean()
s = x.std(ddof=1)

# 검정: H0: mu = 1500, H1: mu < 1500
t_stat = (xbar-1500) / (s/np.sqrt(30))
p_value = stats.t.cdf(t_stat, df=29)
#res = stats.ttest_1samp(x, popmean=1500, alternative="less")

print(f"표본수 n = {n}")
print(f"표본평균 = {xbar:.2f} L, 표준편차 = {s:.2f} L")
print(f"t = {t_stat:.3f},  p-value = {p_value:.4f}")
```

```
표본수 n = 30
표본평균 = 1403.00 L, 표준편차 = 151.57 L
t = -3.505,  p-value = 0.0008
```

두 그룹의 비교

두 그룹의 비교 - A/B검정

- 유의성 검정을 두 개의 그룹을 비교하여 차이가 있는지 판단하는 경우에도 사용할 수 있다.
- 두 그룹을 비교하는 유의성 검정의 또 다른 이름은 A/B 검정(A/B test)이다.
- 두 그룹을 비교하는 경우 “두 그룹이 따르는 확률분포가 같다”, “두 그룹의 평균이 같다”, “두 그룹의 분산이 같다”등을 귀무가설로 생각할 수 있다.
- 이 때 p-값은 귀무가설 하에서 두 그룹의 통계량의 차이가 관측한 값의 차이 또는 그 이상으로 나타날 확률이 되고 이를 주어진 유의수준과 비교한다.

예시

- 어느 제약회사가 새로 개발한 두통약의 효능을 증명하기 위하여 한국에서 1000명의 환자들을 대상으로 임상실험을 수행하여 70%의 치료효과를 보였고, 미국에서 경쟁사의 기존 두통약을 1000명에게 복용하여 60%의 치료효과를 얻었다면 새로운 두통약이 기존의 두통약보다 효과가 좋다고 말 할 수 있는가?
- 두 그룹을 비교하는 실험에서 두 그룹을 구별할 수 있는 의도한 효과를 처리 (treatment)라고 부르며, 위 예시에서는 두통약 투여 (새 두통약, 기존 두통약)가 '처리'가 된다.
- 처리를 제외한 나머지 성질들이 두 그룹에서 동일하여야 공정한 비교가 된다.
- 이를 확보하는 효과적이고 현실적인 수단으로 임의화(randomization)을 시행한다.
- 즉, 실험대상에 처리(treatment)의 배정을 임의로 결정한다.

검정 통계량

- 일반적으로 두 그룹에서 얻는 평균의 차이(또는 차이의 절댓값)을 검정통계량으로 사용한다. $T = \bar{X}_A - \bar{X}_B$.
- 데이터의 분포를 정규분포로 가정하면 평균의 차이도 정규분포를 따름을 알 수 있고 이를 이용하여 p-값을 계산할 수 있다 (two sample t-test)
- 데이터가 충분히 크면 데이터의 분포를 정규분포로 가정하지 않더라도, 중심극한 정리에 의해 평균이 차이가 근사적으로 정규분포를 따른다고 할 수 있다.
- 귀무가설이 맞다면 두 그룹에 차이가 없으므로 두 데이터를 섞어 놓아도 구분할수 없을것이다. 이를 이용한 임의 순열 방법 (permutation method)으로도 p-값을 구할 수 있다.

임의 순열 방법

- 예제를 통해 임의 순열 방법을 소개한다.
- 암 환자 40명을 임의로 20명씩 두 그룹으로 나누어 기존 약과 신약으로 치료한뒤 암세포의 크기를 측정한 데이터가 있다고 하자.

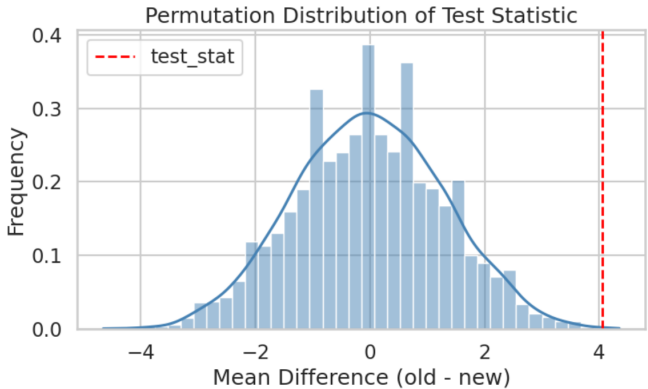
- 이 데이터를 가지고 기존 약과 신약 그룹에서 암세포의 크기에 차이가 유의하게 있는지 유의수준 5%에서 살펴보자
- 이 데이터의 두 그룹의 평균 차이는 $T = \bar{X}_{old} - \bar{X}_{new} = 4.05$이다.
- H_0 : 기존약 그룹과 신약 그룹의 암세포의 크기에 대한 분포는 차이가 없다.
- H_1 : 기존약 그룹 보다 신약 그룹 에서 암세포의 크기가 대체로 더 작다.
- p-값을 구하기 위해서는 귀무가설 하에서 검정통계량 T의 분포를 구해야 한다.

- 기존 약과 신약에서의 암세포의 크기의 분포가 같다면 두 데이터를 섞어놓아도 구분할 수 없을것이므로 두 그룹의 데이터를 임의로 섞어서 다시 두 그룹으로 나눈 후 평균의 차이를 구하는 과정을 반복하여 평균의 차이에 대한 경험적 분포(empirical distribution)을 구한다. 즉, 다음과 같은 과정을 진행한다.

(1) 총 B번 a,b과정을 반복한다.
   a 총 40명의 데이터에 대해 기존약과 신약을 임의로 재 배정한다.
   b 재 배정된 데이터를 기반으로 $T = \bar{X}_{old} - \bar{X}_{new}$를 계산하여 값을 저장한다.

(2) 저장된 B개의 값을 이용하여 T의 분포를 구한다.

(3) 데이터로부터 구한 관측값인 6.05와 비교하여 p-값을 구한다.



- 임의순열방법을 통해 구한 p-값에 대한 근사값이 유의수준 5%보다 작으므로 귀무가설을 기각한다.
- 따라서 신약은 암세포의 크기를 줄이는데 기존약보다 효과적이라고 할 수 있다.

The End