Chapter 03. 통계적 실험과 유의성 검정 - 유병욱

> 출처 : https://liam427.github.io/mathmatics/PracticalStatistics_03/
> 유병욱님 정리자료
>
> > **Practical Statistics for Data Scientists - 통계적 실험과 유의성 검정**
>
> Updated: @2021년 9월 10일

3. 통계적 실험과 유의성검정

- 실험설계는 어떤 가설을 확인하거나 기각하기 위한 목표를 갖고 있다. 특히나 데이터 과학자들은 종종 사용자 인터페이스나 제품 마케팅 실험과 같이 지속적으로 어떤 실험을 수행해야 하는 상황에 있다.
- 추론이라는 용어는 제한된 데이터로 주어진 실험 결과를 더 큰 과정 또는 모집단에 적용하려는 의도를 반영한다.

3.1 A/B검정

- **A/B검정**은 두 가지 처리 방법, 제품, 절차 중 어느 쪽이 다른 쪽보다 더 우월하다는 것을 입증하기 위해 실험군을 두 그룹으로 나누어 진행하는 실험이다. 종종 두 가지 처리 방법 중 하나는 기준이 되는 기존 방법이거나 아예 아무런 처리도 적용하지 않는 방법이 된다. 이를 대조군이라고 하며, 새로운 처리 방법을 적용하는 것이 대조군보다 더 낫다는 것이 일반적인 가설이 된다.
- **A/B검정**의 예
  - 종자 발아가 어디에서 더 잘되는지 알아보기 위해 두 가지 토양 처리를 검정한다.
  - 암을 더 효과적으로 억제하는 두 가지 치료법을 검정한다.

- 두 가지 가격을 검정하여 더 많은 순이익을 산출하는 쪽을 결정한다.
- 두 개의 인터넷 뉴스 제목을 검정하여 더 많은 클릭을 생성하는 쪽을 결정한다.
- 두 개의 인터넷 광고를 검정하여 어느 것이 더 높은 전환율을 얻을지 판단한다.

3.1.1 대조군은 왜 필요할까?

- 대조군이 없다면 '모든 다른 것들은 동일하다.'는 보장이 없으며 어떤 차이가 처리(또는 우연)때문인지 확신할 수 없다. 대조군이 아닌 단순히 이전의 경험과 비교할 경우, 처리 이외의 다른 요소가 다를 수도 있기 때문이다.
- 여러 행동 유형과 관련된 지표들이 수집 대상이 될 수 있지만, 실험이 결국 처리 A와 처리 B 사이의 결정으로 이어질 경우, 단일 지표 또는 검정통계량(처리 효과를 측정하기 위한 지표)을 사전에 미리 정해놓아야 한다. 실험을 수행한 뒤 나중에 검정통계량을 선택한다면 연구자 편향이라는 함정에 빠지게 된다.

3.2 가설검정

- 가설검정 혹은 유의성검정은 지금까지 발표된 대부분의 연구 논문에 등장하는 전통적인 통계분석 방법이다. 목적은 관찰된 효과가 우연에 의한 것인지 여부를 알아내는 것이다.
- 적절하게 설계된 A/B검정에서는, A와 B 사이의 관찰된 차이가 다음 원인들로 설명될 수 있도록 A와 B에 대한 데이터를 수집한다.
  - 우연한 대상 선정
  - A와 B의 진정한 차이

3.2.1 귀무가설 h0

- 그룹들이 보이는 결과는 서로 동일하며, 그룹 간의 차이는 우연에 의한 결과라는 것을 기본 가정으로 설정한다. 이 기본 가정을 **귀무가설**이라고 부른다. 결국, 귀무가설이 틀렸다는 것을 입증해서, A 그룹과 B 그룹 간의 차이가 우연이 아니라는 것을 보여주는 것이 목적이다.

3.2.2 대립가설 h1

- 가설검정의 예
  - 귀무가설 : 그룹 A와 그룹 B의 평균에는 차이가 없다. vs 대립가설 : A는 B와 다르다
  - 귀무가설 : $A \leq B$ vs 대립가설 : $A > B$

- 귀무가설과 대립가설이 모든 가능성을 설명할 수 있어야 한다.
  귀무가설의 본질은 가설검정의 구조를 결정한다.

3.2.3 일원/이원 가설검정

- 방향성을 고려한(단방향) 대립가설이 필요하다.(B는 A보다 낫다.) 이 경우 일원 가설검정을 사용한다. 즉, 우연에 의한 극단적인 결과에 대해 한 방향만을 고려하여 p 값을 계산한다는 의미이다.
- 어느 쪽으로도 속지 않도록 가설검정을 원한다면 대립가설은 양방향(A는 B와 다르며 더 크거나 더 작을 수 있음)이 된다. 이 경우 이원가설을 사용한다. 우연에 의한 극단적인 결과가 양쪽에 나타날 p 값을 계산한다는 것을 의미한다.

- **귀무가설**은 우리가 관찰한 어떤 효과가 특별한 것이 아니고 우연에 의해 발생한 것이라는 개념을 구체화하는 일종의 논리적 구조이다.
- **가설검정**은 귀무가설이 사실이라고 가정하고, 영모형을 생성하여 관찰한 효과가 해당 모델로부터 합리적으로 나올 수 있는 결과인지 검증하는 것이다.

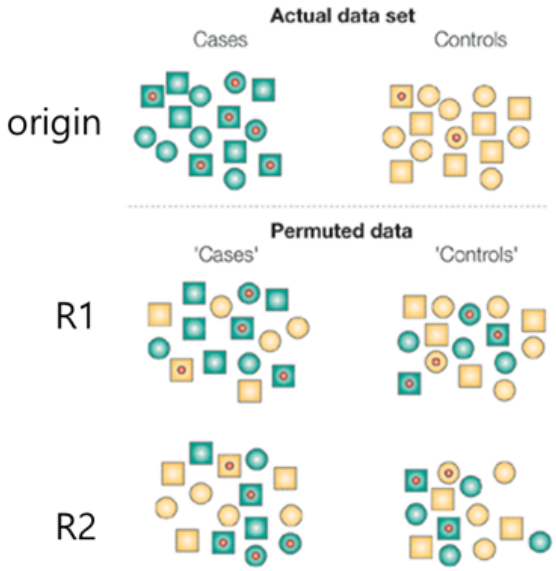
재표본추출

- **재표본추출**이란 랜덤한 변동성을 알아보자는 일반적인 목표를 가지고, 관찰된 데이터의 값에서 표본을 반복적으로 추출하는 것을 의미한다.

3.3.1 순열검정

- 순열과정에는 두 개 이상의 표본이 관여되며 이들은 통상적으로 A/B 또는 기타 가설검정을 위해 사용되는 그룹들이다.
  - 순열
- 순열의 절차
  - 여러 그룹의 결과를 단일 데이터 집합으로 결합한다.
  - 결합된 데이터를 잘 섞은 후, 그룹 A와 동일한 크기의 표본을 무작위(비복원)로 추출한다.
  - 나머지 데이터에서 그룹 B와 동일한 크기의 샘플을 무작위(비복원)로 추출한다.
  - C, D 등의 그룹에 대해서도 동일한 작업을 수행한다. 이제 원본 표본의 크기를 반영하는 재표본을 수집했다.
  - 원래 샘플에 대해 구한 통계량 또는 추정치가 무엇이었든간에 지금 추출한 재표본에 대해 모두 다시 계산하고 기록한다.

- 앞선 단계들을 R번 반복하여 검정통계량의 순열분포를 얻는다.



- 관찰된 차이가 순열로 보이는 차이의 집한 안에 잘 들어 있다면, 우리는 어떤것도 증명할 수 없다. 즉, 관찰된 차이가 우연히 일어날 수 있는 범위 안에 있다는 말이다. 하지만 관찰된 차이가 대부분의 순열분포 바깥에 있다면, 우리는 이것은 우연 때문이 아니라고 결론을 내릴 수 있다. 전문적인 표현으로, 이 차이는 통계적으로 유의미하다.

3.3.3 전체 및 부트스트랩 순열검정

- 랜덤 셔플링 절차를 '임의순열검정' 또는 '임의화 검정'이라고 부른다.
  - 전체순열검정
  - 부트스트랩 순열검정
- **전체순열검정** : 데이터를 무작위로 섞고 나누는 과정에서 나눌 수 있는 *모든 가능한 조합*을 찾는다. 따라서 샘플 크기가 비교적 작을 때만 실용적이다. 셔플링을 많이 반복할수록, 임의순열검정 결과는 전체순열검정 결과와 유사하게 근접한다. '유의미하다'라는 결론이 아닌 더 정확한 결론을 보장하는 통계적 속성이 있어서 *'정확검정'*이라고도 한다.
- **부트스트랩 순열검정** : 무작위 순열검정의 2,3단계에서 비복원으로 하던 것을 복원 추출로 수행한다. 이는 리샘플링 과정에서 모집단 개체를 선택할 때, 개체가 다시 그룹에 할당될 때에도 임의성을 보장한다. 하지만, 이를 구별하는 일이 복잡하고, *데이터 과학에서 별로 실용적이지 않다.*

3.4 통계적 유의성과 p값

- 통계적 유의성이란, 통계학자가 자신의 실험 결과가 우연히 일어난 것인지 아니면 우연히 일어날 수 없는 극단적인 것인지를 판단하는 방법이다. 결과가 우연히 벌어질 수 있는 변동성의 바깥에 존재한다면 우리는 이것을 통계적으로 유의하다고 말한다.

3.4.1 p 값

- **p 값**과 같이 통계적 유의성을 정확히 측정하기 위한 지표가 필요하다. 확률모형이 관측된 결과보다 더 극단적인 결과를 생성하는 빈도라고 할 수 있다.

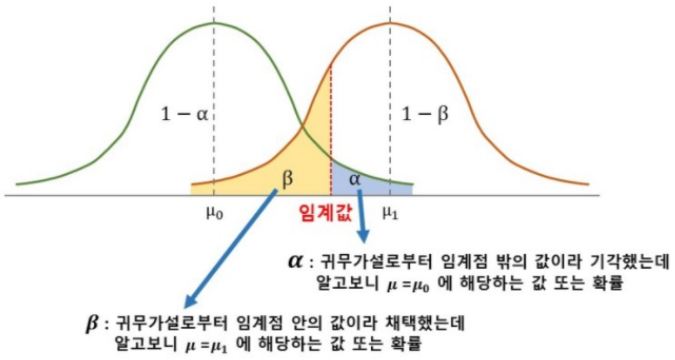
3.4.2 유의수준

- 우현히 얻은(귀무가설) 결과의 5%보다 더 극단적인 결과와 같이 어떤 임곗값(5%)을 미리 지정하는 것을 선호한다. 이 임곗값을 보통 유의수준(α)이라고 한다. 많이 사용되는 유의수준 5%와 1%이다.

  α

3.4.3 제 1종과 제 2종 오류

		검정의 결과	
		H_0 기각하지 않음	H_0 기각
실제	H_0 참	올바른 판단	제1종 오류
	H_1 참	제2종 오류	올바른 판단



- 보통은 1종 오류를 최소화하도록 가설을 설계한다.

> 어떤 비료 한 바구니는 7.25 kg라고 홍보되고 있지만, 실제로는 정규분포를 따르며 평균이 7.4kg, 표준편차가 0.15kg이다.비료회사는 기계를 설치하여 비료 양의 평균이 변화하는지 확인한다.

> H0 : μ = 7.4kgH1 : μ ≠ 7.4kg

> 비료 50바구니를 임의추출하고 그 결과 표본평균는 7.36kg, 표준편차는 0.12kg이다.p값을 구했더니 0.02이다.

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{7.36 - 7.4}{0.12/\sqrt{50}}$$

> 유의수준 α = 0.05 를 이용하여 내린 결론은?=> p < α 이므로 귀무가설 기각, 대립가설 수용=> 바구니들에 채워진 양의 평균이 7.4kg이 아닙니다.유의 수준이 α = 0.01이라면 결론은?=> p > α 이므로 귀무가설 기각 실패!

3.5 t 검정

- 평균 검정이란 단일 or 독립 집단 사이의 가설 검증을 위한 수단이다.
  집단의 평균의 차를 비교하기 위해 **수치형 변수**를 사용한다.
  검정통계량(평균검정 : t값)과 p-value를 계산하여 신뢰구간을 만족하는지 확인하고, 가설의 채택 여부를 결정한다.
- 평균검정에는 z-검정, t-검정, 분산분석으로 나뉜다.
  - z-검정과 t-검정은 비교 집단이 2개 이하일 경우, 분산분석은 비교 집단이 3개 이상일 경우 사용한다.
  - z-검정은 모분산을 알고 있을 때만 사용이 가능한 반면, t-검정은 모분산을 모를 때도 사용이 가능하다.

3.5.1 1-sample T-test

- 기존에 알려져 있는 평균 값이 맞는지를 확인하기 위한 검정방법
  - 일반적으로 모집단의 평균이 특정 값으로 알려져있는 경우, 실제로 모집단의 평균이 특정 값과 같은지에 대해 가설을 세우고 검정한다.

```
import numpy as np
from scipy import stats

# 초기값 고정
np.random.seed(333)

# 평균 = 160, 표준편차는 3으로 하는 50개의 랜덤값을 만들자.
heights = [160 + np.random.normal(0, 3) for _ in range(50)]
```

```
# t-test : stats.ttest_1samp
tTest = stats.ttest_1samp(heights, 155)

# print result
print("The T-statistic is %.3f and the p-value is %.3f" % tTest)
```

```
The T-statistic is 11.173 and the p-value is 0.000
```

- p-value가 0으로, 기각역을 p < 0.05로 설정했을때 귀무가설을 기각한다. 즉, 학생들의 실제 평균 키가 160일 때, 위와 같은 표본을 얻을 확률이 0으로, 학생들의 평균키는 160이 아니라고 할 수 있다.

3.5.2 독립표본 t-test(Unpaired T-test)

- 독립된 두 집단의 평균 차이를 검정하는 기법이다. 반드시 서로 무관한 독립된 두 집단을 사용해야한다. 독립표본 t-검정의 경우, 등분산 여부에 따라 결과값이 달라지기 때문에 독립표본 t-검정을 시행하기 전에 등분산검정을 시행한 후 그 결과에 따라 독립표본 t-검정을 시행한다.
  - 등분산검정 : 두 모집단에서 추출한 표본의 분산이 같은 것을 말한다.
  - 분산이 다르다 ⇒ 편차가 다르다 ⇒ 분산이 같아야 평균이 같은지 비교하는게 의미가 있다는 뜻

3.5.3 쌍체표본 t-검정(Paired T-test)

- 동일한 항목, 사람, 물건에 대한 측정 값이 두개인 경우 그 차이를 비교하기 위해 사용하는 검정방법. 이 때 분석 대상의 **표본은 반드시 대응**되어야 한다.
  (만약 대응되지 않으면 결측값이 있다는 의미이므로 결측값을 처리한 후 분석 진행) 또한, 대응표본은 시간의 개념이 있어 꼭 독립된 두 집단이 아니어도 된다.

3.6 다중검정

- 다중검정은 검정군이 3가지 이상인 경우에 5%의 유의수준으로 유의미하다고 말할 수 있을까? 로부터 시작된다.

> 예를 들어 실험군이 1000개가 있다.실험군 1이 나머지 999개의 실험군보다 p<0.05 로 유미하게 차이가 난다고 하면,실험군 1이 2와 유의미한 차이가 나지 않지만, 차이가 난다고 잘못 결과를 내릴 확률(1종오류)은 5%이다.실험군 1이 3와 유의미한 차이가 나지 않지만, 차이가

> 난다고 잘못 결과를 내릴 확률(1종오류)은 5%이다.실험군 1이 4와 유의미한 차이가 나지 않지만, 차이가 난다고 잘못 결과를 내릴 확률(1종오류)은 5%이다.…실험군 1이 999와 유의미한 차이가 나지 않지만, 차이가 난다고 잘못 결과를 내릴 확률(1종오류)은 5%이다.이 확률을 모두 곱하면 999 * 0.05 = 약50이다. 즉, 50개의 검정이 잘못되었을 수 있다.이런 문제를 다중 검정 비교의 문제 라고 한다. 따라서 다중 검정인 경우에는 p-value설정 만으로 끝나는 것이 아니라 사후보정을 해주어야 한다.

3.6.1 다중 검정 보정 방법

- 본페로니 보정 방법
  - FWER(Family Wise 1 Error Rate) : 한 연구에서 적어도 한 개의 잘못된 결론이 나올 수 있는 확률
  - FWER을 통제하기 위한 방법 중 본페로니 보정 방법이 있다. test의 수가 n이고, FWER을 0.05로 통제할 때, 개별 테스트의 유의 수준을 α/n으로 설정한다. 모든 검정이 실제로 연관성이 없을 때는 n이 클수록 대략적으로 만족시킨다. 하지만, 이 방법은 너무 보수적(귀무가설을 웬만하면 채택)이기 때문에, FP는 줄일 수 있지만 FN은 많아진다.

    α/n

  -

실제 상황 (ground truth)	예측 결과 (predict result)	
	Positive	Negative
Positive	TP(true positive) **옳은 검출**	FN(false negative) **검출되어야 할 것이 검출되지 않았음**
Negative	FP(false positive) **틀린 검출**	TN(true negative) **검출되지 말아야 할 것이 검출되지 않았음**

- FDR(False Discovery Rate)
  - FDR을 다중검정에서 사용할 때의 의미는 FP, FN으로의 집중이 아니라 내가 귀무가설을 기각한 검정 중 틀린 것의 비율을 줄이자는 것이다.

$$FDR = \frac{false_{positive}}{true_{positive} + false_{positive}}$$

3.7 자유도Permalink

- 자유도란 통계적 추정을 할 때 표본자료 중 모집단에 대한 정보를 주는 독립적인 자료의 수를 말한다.

	당뇨	정상	전체
고혈압	*a*	*b*	20
정상	*c*	*d*	80
전체	25	75	100

자유도는? 1

- 2차원 행렬에서도 자유도를 알 수 있다. a, b, c, d의 값 중 하나만 정해지면 다른 값들이 모두 결정되기 때문에 자유도는 1 이다.

	당뇨	내당능장애	정상	전체
고혈압	*a*	*b*	*c*	20
정상	*d*	*e*	*f*	80
전체	25	25	50	100

자유도는? 2

- 3x2 행렬에서는 2개의 값이 정해지면 나머지 값을 모두 채울 수 있다.
  - 이렇게 n x m의 행령은 (n-1)X(m-1)의 자유도를 따른다.
- 자유도가 필요한 이유는?
  - 자유도는 모분산을 모르기 때문에 필요하다. 모집단에서 표본을 추출하면 표본의 평균은 모집단의 평균에 대해 클 수도 있고 적을수도 있으나 그 가능성은 공평하다.(=불편 추정) 그러나 표본의 분산은 모집단의 분산보다 항상 작아지는 경향을 보인다.(= 편향의 경향성을 띔) 따라서 표본의 분산을 모집단의 분산에 근사해지게 하는 비율을 찾게 되었다. 이 비율은 n/(n-1) 이고, 표본의 분산에 이 비율을 곱하면 모집단의 분산에 근사하게 된다. 그런데 분산의 원래 계산식에 있는 분모의 n이 약분되기 때문에 (n-1)만 남게 된다. 결국 표본의 분산을 구할 때, n 대신 n-1을 나누면 표본의 분산을 모집단의 분산에 근사해지게 할 수 있다. 따라서 자유도는 표본의 평균을 구할 때는 사용되지 않고 표본의 분산을 구할 때만 사용된다.
- 표본의 분산은 왜 모집단의 분산보다 작을까?

    - 분산은 제곱한 값들로 이루어져있기 때문이다. 따라서 집단의 크기가 큰 모집단의 분산은 당연히 집단의 크기가 비교적 작은 표본의 분산보다 클 수 밖에 없다. 하지만, 표본의 개수가 커짐에 따라 표본분산과 모분산과의 차이가 작아진다. 따라서 30개 이상 또는 그 이상의 대표본에 대해서는 표본의 분산을 구할 때 자유도(n-1)를 고려하지 않아도 된다.

3.8 분산분석Permalink

- A/B검정은 두 개의 그룹을 비교하였다. 그렇다면 A, B, C, D 그룹의 데이터를 비교한다고 가정하자. 여러 그룹간의 통계적으로 유의미한 차이를 검정하는 통계적 절차를 분산분석(ANOVA)이라고 한다.
- 분산분석의 조건
    - 정규성 : 각각의 그룹에서 변인은 정규분포이다.
    - 분산의 동질성 : y의 모집단 분산은 각각의 모집단에서 동일하다.
    - 관찰의 독립성 : 각각의 모집단에서 크기가 각각인 표본들이 독립적으로 표집되어 있다.

ANOVA

3.8.1 일원 분산분석Permalink

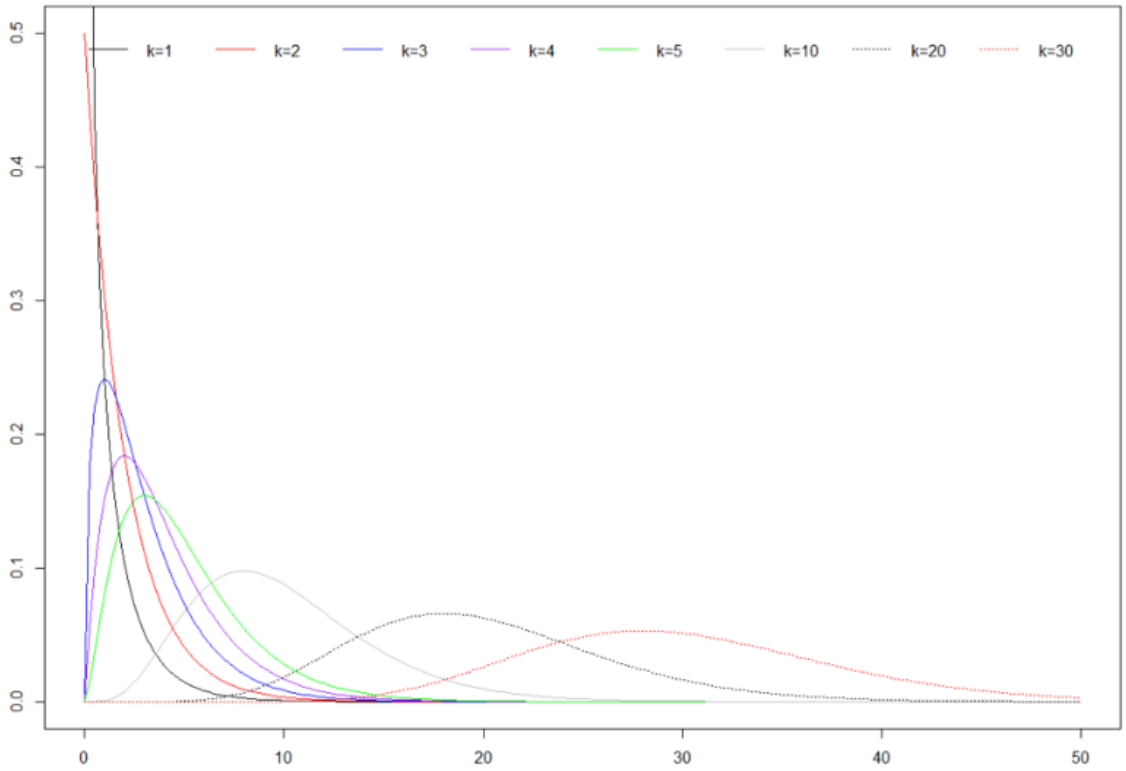
- 종속 변인과 독립 변인의 집단이 1개인 경우이다. 한 가지 변수의 변화가 결과 변수에 어떤 영향을 미치는지 확인할 수 있다.

3.8.2 이원 분산분석Permalink

- 독립변인의 수가 두 개 이상일 때 집단 간 차이가 유의한지 검증하는데 사용된다.
- 만약 페이지를 접속하는 수가 주말/평일에 따라 많은 영향을 받는다고 한다면(페이지1 주말, 페이지1평일, 페이지2주말, 페이지2평일 등)의 데이터가 생성된다. 이 때 필요한 것이 이원분산분석이다. 이는 상호작용 효과를 확인하는 식으로 일원분산분석과 방식은 비슷하다.
- 분산분석의 한계점
    - 전체 그룹 간 평균값 차이가 통계적 의미가 있는지 판단하는데 유용하지만, 어느 그룹의 평균값이 의미가 있는지는 알기 어렵다. 따라서 추가적인 사후 분석이 필요하다.

3.9 카이제곱검정Permalink

- 정규분포를 따르는 모집단에서 크기가 n인 표본을 무작위로 반복 추출한다. 이 때 각 표본의 분산들이 카이제곱 분포를 따른다고 한다. 자유도가 커질수록 정규분포에 가까워지며 다음과 같이 나타난다.



- 카이제곱 검정방법
  - 독립성 검정 : 두 변수는 서로 연관성이 있는가?
  - 적합성 검정 : 실제 표본이 내가 생각하는 분포와 같은가?
  - 동일성 검정 : 두 집단의 분포가 동일한가?

3.9.1 일원 카이제곱 검정 → 적합성 검정Permalink

```
import pandas as pd

# 관찰빈도
xo = [324, 78, 261]
#기대빈도
xe = [371, 80, 212]
# df
xc = pd.DataFrame([xo, xe],
                  columns=['a', 'b', 'c'],
                  index=['Obs', 'Exp'])

xc
```

Aa 제목	속성	# a	# b	# c
제목 없음	**Obs**	324	78	261
제목 없음	**Exp**	371	80	212

```
from scipy.stats import chisquare

result = chisquare(xo, f_exp=xe)
result
```

```
Power_divergenceResult(statistic=17.329649595687332, pvalue=0.00017254977751013492)
```

- p-value가 0.000172549로 유의수준 0.05보다 아주 작으므로 귀무가설을 기각하고, 대립가설을 채택한다. 즉, 관찰빈도와 기대빈도는 다르다.

이원 카이제곱 검정 → 독립성, 동질성 검정Permalink

```
xf = [269, 83, 215]
xm = [155, 57, 181]
x = pd.DataFrame([xf, xm],
                 columns = ['item1','item2','item3'],
                 index = ['Female','Male'])
x
```

Aa 제목	속성	# item1	# item2	# item3
제목 없음	**Female**	269	83	215
제목 없음	**Male**	155	57	181

- 독립성 검정
  - 귀무가설 : 성별과 아이템 품목 판매량은 관계가 있다.
  - 대립가설 : 성별과 아이템 품목 판매량은 관계가 없다.

```
from scipy.stats import chi2_contingency
chi_2, p, dof, expected = chi2_contingency([xf, xm])
msg = 'Test Statistic: {}\np-value: {}\nDegree of Freedom: {}'
print(msg.format(chi_2, p, dof))
print(expected)
```

```
Test Statistic: 7.094264414804222
p-value: 0.028807134195296135
Degree of Freedom: 2
[[250.425   82.6875 233.8875]
 [173.575   57.3125 162.1125]]
```

- p-value는 0.02881으로 유의수준 0.05보다 작은 값이므로 귀무가설을 기각한다. 따라서 성별과 아이템 품목 판매량은 관계가 없다.

3.10 멀티암드 밴딧 알고리즘Permalink

- 실험설계에 대한 전통적인 통계적 접근 방식보다 명시적인 최적화와 좀 더 빠른 의사 결정을 가능하게 하며, 여러 테스트, 특히 웹 테스트를 위해 사용된다.
- 슬롯머신의 팔(=밴딧)을 당기면 돈을 얻게 된다. 만약, 슬롯머신에 둘 이상의 손잡이가 달려있고 각 손잡이가 다른 속도로 돈을 지불한다면 우리는 많은 상금이 나오는 손잡이를 빨리 확인하고자 할 것이다. 다음과 같이 가정해보자.

> 손잡이 A : 50번 중 10번 승리손잡이 B : 50번 중 2번 승리손잡이 C : 50번 중 4번 승리

- 그러면 우리는 A가 최고의 손잡이인 것 처럼 보인다. 하지만 사실 B,C가 더 좋다면 우리는 A를 당기기 때문에 B,C를 놓치게 된다. 따라서 하이브리드 접근 방식을 취하여 A의 우위를 적극 활용해보자. C를 잡아당길 기회를 A에게 더 준다. 그러다가 A가 나빠지기 시작하면 기회를 다시 C에게 돌린다. 그 중 하나가 A보다 우수하고 이것이 초기 실험에서 감춰졌었다면 사실을 밝힐 수 있게된다.

3.10.1 웹 테스트 적용Permalink

- 이것을 웹 테스트에 적용한다면? 여러 개의 손잡이 대신에 제안/헤드라인/색상을 테스트 할 수 있다. 고객은 클릭/ 클릭안함의 결정을 한다. 처음에는 무작위로 균등하지만 하나가 좋은 결과를 내기 시작하면 더 자주 표시될 수 있도록 한다.
- 그렇다면 잡아당기는 비율을 언제 어떻게 수정해야할까?
  - 엡실론 - 그리디 알고리즘 (epsilon - greedy algorithm)
    - 1 : A/B 검정
    - 0 : 탐욕알고리즘 = 더 이상의 실험 없이, 피실험자 웹 방문자들을 지금까지 알려진 가장 좋은 제안에 할당한다.
  - 톰슨의 샘플링 (Thompson’s sampling)
    - 표본을 추출하여 최고의 손잡이를 선택할 확률을 최대화한다.

    - 베이지언 방식 : 베타분포를 사용하여 수익의 일부 사전 분포를 가정한 뒤, 각 정보가 누적되면서 업데이터되어 다음번에 최고 손잡이를 선택할 확률을 효과적으로 최적화할 수 있다

MAB

3.11 검정력과 표본크기Permalink

- 웹 테스트를 수행할 경우 실행 시간은 어떻게 결정할까? 웹 테스트에 대한 수많은 관련 자료들을 인터넷에서 쉽게 찾을 수 있다. 하지만 모든 경우에 딱 맞는 일반적인 방법은 없고, 다만 원하는 달성 목표에 따라 조절해야 한다.
- 표본크기에 대한 고려는 '가설검정이 실제로 처리 A와 B의 차이를 밝혀낼 수 있을까?'라는 질문과 바로 연결된다. 가설검정의 결과라고 할 수 있는 P 값은 A와 B사이에 실제 차이가 있는지에 따라 달라진다. 물론 실험에서 누가 어떤 그룹에 속하느냐는 선택의 운에 따라 결과가 달라질 수도 있다. 그렇다 하더라도 실제 차이가 크면 클수록, 그것을 밝혀낼 가능성도 따라서 커질 것이고, 그 차이가 작을수록 더 많은 데이터가 필요하다는 생각에는 모두 동의할 수 있다. 야구에서 3할 5푼 타자와 2할 타자를 구분하기 위해 많은 타석이 필요하지는 않다. 하지만 3할 타자와 2할 8푼 타자를 구분하기 위해서는 더 많은 타석 정보가 필요할 것이다.

3.11.1 검정력Permalink

- 검정력이란 바로 특정 표본 조건에서 특정한 효과크기를 알아낼 수 있는 확률을 의미한다. 예를 들어 25타석에서 3할 3푼 타자와 2할 타자를 구분할 수 있을 확률이 0.75라고 말할 수 있다. 여기서 효과크기란 바로 1할 3푼의 타율 차이 (0.130)를 의미한다.
- 그리고 '알아낸다'는 것은 가설검정을 통해 차이가 없을 것이라는 영가설을 기각하고 실제 효과가 있다고 결론을 내리는 것을 의미한다. 다시 말해 두 타자를 대상으로한 25타석(N=25) 실험은 0.130의 효과크기에 대해 0.75 혹은 75%의 검정력을 가진다고 볼 수 있다.
- 몇 가지 '움직이는 부분'이 있다. 가설검정에서 표본 변이, 효과크기, 표본크기, 유의수준 등을 특정하는데 필요한 수많은 통계적 가설과 수식에 말려들기 십상이다. 실제로 검정력을 계산하기 위한 특별한 목적의 통계 소프트웨어가 있다. 대부분의 데이터 과학자들은 검정력을 구하기 위해 형식적인 절차를 모두 지킬 필요는 거의 없다. 하지만 A/B 검정을 위해 데이터를 수집하고 처리하는데 비용이 발생하는 경우, 가끔 사용해야 할 수도 있다. 이럴 경우, 데이터 수집을 위해 대충 얼마의 비용이 발생할지 안다면 데이터를 수집하고도 결론을 내리지 못하는 상황을 피할 수 있을 것이다. 여기 꽤 직관적인 방법 하나를 소개한다.

1. 최대한 결과 데이터가 비슷하게 나올 수 있는 가상의 데이터를 생각해보자. 예를 들면 2할 타자를 위해 20개의 1과 80개의 0이 들어 있는 상자를 생각한다든지, 아니면 웹 페이지 방문 시간을 관측한 자료가 담겨 있는 상자를 생각할 수 있다.
2. 첫 표본에서 원하는 효과크기를 더해서 두 번째 표본을 만든다. 예를 들면 33개의 1과 67개의 0을 가진 두 번째 상자, 혹은 각 초기 방문시간에 25초를 더한 두 번째 상자를 만들 수 있다.
3. 각 상자에서 크기 N인 부트스트렙 표본을 추출한다.
4. 두 부트스트랩 표본에 대해서 순열 가설검정을 진행한다.
5. 3~4단계를 여러 번 반복한 후, 얼마나 자주 유의미한 차이가 발견되는지 알아본다. 이 확률이 바로 검정력 추정치다.

3.11.2 표본크기Permalink

- 검정력 계산의 주된 용도는 표본크기가 어느 정도 필요한가를 추정하는 것이다.
- 예를 들면 기존 광고와 새로운 광고를 비교하기 위해 클릭률을 조사한다고 가정하자. 이 조사를 위해 얼마나 많은 클릭 수를 수집해야 할까? 50% 정도의 큰 차이에만 관심이 있다면 상대적으로 적은 수의 표본으로도 목표를 이룰 수 있을 것이다. 하지만 그것보다 훨씬 작은 차이에도 관심이 있다면 훨씬 큰 표본이 필요하다. 이런 식으로 새 광고가 기존 광고에 비해 얼마큼 더 효과적이야 하는지, 어느 정도가 아니면 기존 광고로 계속 갈지에 대한 기준을 설정하는 것이 표준적인 접근법이다. 이러한 목표, 즉 '효과크기'가 표본크기를 좌우한다.
- 예를 들면 현재 클릭률이 약 1.1% 수준인데 여기서 10% 증가한 1.21%를 원한다고 가정하자. 이때 우리는 두 상자, 1.1%의 1이 들어 있는 상자 A와 1.21%의 1이 들어 있는 상자 B가 있다고 생각할 수 있다. 먼저 각 상자에서 300개씩 뽑는다고 하자. 결과가 다음과 같다고 가정하자.
    - 상자 A : 3개의 1
    - 상자 B : 5개의 1
- 어떤 가설검정을 해도 이 차이가 유의미하지 않다고 나올 것이라고 쉽게 눈치챘을 것이다. 이 표본크기와 효과크기의 조합은 가설검정을 통해 이 차이를 보이기에는 너무 작다.
- 따라서 이번에는 표본크기를 증가시켜 2,000개의 첫인사을 알아보자. 그리고 더 큰 효과크기를 생각해보자.
- 다시 클릭률은 여전히 1.1% 수준이라고 가정한다. 대신 50% 증가한 1.65를 원한다고 생각해보자. 아까와 마찬가지로 두 상자, 1.1%의 1이 들어 있는 상자 A와 1.65%의 1이

들어 있는 상자 B가 있다고 생각할 수 있다. 이제 각 상자에서 2,000개를 뽑는다. 이렇게 뽑은 결과가 다음과 같다고 하자.

  - 상자 A : 19개의 1
  - 상자 B : 34개의 1

- 하지만 유의성 검정을 해도 이 차이(34-19)가 여전히 '유의미하지 않다'라고 결론 날 것이다. 검정력을 계산하기 위해서는 이러한 과정을 여러 번 반복해야 한다. 아니면 검정력 계산을 지원하는 소프트웨어를 사용할 수도 있다. 하지만 앞 예제를 통해 알 수 있듯이 50% 정도의 효과를 알기 위해선 수천개 이상의 광고 첫인상 정보가 필요할 것이다.
- 요약하면, 검정력 혹은 필요한 표본크기의 계산과 관련한 다음 4 가지 중요한 요소들이 있다.
  - 표본크기
  - 탐지하고자 하는 효과크기
  - 가설검정을 위한 유의수준
  - 검정력
- 이 중 3가지를 정하면 나머지 하나를 알 수 있다. 가장 일반적으로, 표본크기를 알고 싶을 경우가 많다.