**표본평균과 실제 모집단의 M와 얼마나 가까운 가?**

신뢰도와 신뢰구간

* 신뢰도를 높이기 위해서 신뢰구간을 무한히 크게 하면 정교하지 않음

표준오차(Standard error) 와 표준편차(Standard error)

* 표준오차 : 표본평균의 분산^1/2(즉, 추정 통계량의 표준편차)
* 표준편차 : 모집단에서 도출된 분산^½

T통계량

* 모집단의 분산을 실제로 모를때, 표본평균의 분산과 유사할 것이라고 가정(N이 충분히 크다면!)
* t분포 t(n-1)

가설 검정

* 대립가설이 맞다고 검증하는 것이 아니라, 귀무가설에 대한 검증을 함.
* 귀무가설 ⇒ 기각 or 기각할 수 없음(채택과는 다름)
* 귀무가설이 기각 됐다는 것은, 대립가설이 맞다가 아니라 귀무가설이 기각되었기 때문에 대립가설이 지지받는다는 것임(통계적으로 엄밀히는 그렇다)
* 실험을 딱 1번 했을때 나오는 값이 일반적으로 나오는 값은 아니라면?
* 실험을 딱 1번해서 나온 값이 나올 확률이 말도 안되게 적은 확률이었다면?

P value -> P value를 보고 판단하는 기준값을 유의수준이라고 함

* 귀무가설이 맞다고 가정했을 때, 내가 관측하는 값이 정말 극단적으로 더 이상할 확률
* P value가 작으면 더 극단적으로 일어날 확률이 적은거니깐, 이미 충분히 극단적임
  + 귀무가설이 잘못되서 실험 1번 했지만 충분히 극단적으로 나온 것
* P value가 크면 더 극단적으로 일어날 확률이 많으니깐, 이미 충분히 정상적일 수 있음
* 전통적으로 유의수준을 5% 사용했지만, 필드마다 1%, 0.1%, 0.01%일 수도 있음.
* 기본적으로 데이터의 양이 많아질수록 유의수준은 낮아져야 함

통계는 기본적으로 여러 값에 대한 것. 하나의 값에 대한 검정은 통계를 쓸 필요가 없음.

평균에 대해서 많이 보는 이유는, 가장 간단하고 이론적으로 탄탄한 것이기 때문

* 원래는 평균 이외에도 많은 통계량에 대해 검정이 가능함(컴퓨터를 활용하여)
* 시뮬레이션을 통해 수치적으로 검정 가능함(수식값이나 이론으로는 여전히 어려움)

**독립된 두 집단의 모평균 검정(엄밀히는 두 변수간의 독립성 검정임)**

두 랜덤 변수의 데이터 타입에 따라 관계 검정이 달라짐

⇒ 데이터를 보는 관점이 최근에는 문제를 인식하고 데이터를 모으는 것이 아니라, 데이터는 이미 모여있는 상태에서 분석함(키에 대한 두 그룹이 아니라, 키 컬럼하나 & 성별 컬럼하나)

|  |  |  |
| --- | --- | --- |
| X/Y | NUMERIC | Category |
| NUMERIC | Correlation | T-test/ANOVA |
| Category | T-test/ANOVA | Chi-square |

ex) 성별(남,녀-범주) 와 키(숫자) => T-test/ANOVA

가장 근본적으로 관계 검정은 두 변수간의 독립성 검정이다.

Pr(X=x) = Pr(X | Y=F) = Pr( X | Y=M) 이면 독립

여성과 남성의 분포를 같은지에 대한 검정을 하는 것이 정확하나, 이는 어려움

⇒ 여성집단의 평균과 남성집단의 평균과 같으냐 다르냐로 약한 조건으로 검정

여성과 남성의 평균이 다르면, 키와 성별(X와 Y)는 밀접한 관련이 있다고 볼 수 있음.

**평균이 다르면 관련이 있는 것임.**

**평균이 같아도(T-test로는 못 찾음, 관련이 없다고 나오니깐) 관련이 있는 경우가 있을 수도 있음 ⇒ 다른 검정방법을 써야함**

**⇒ 근본적인 관련성을 검정하는 것은 매우 어려워서, 평균의 차이로 간단하게 검정하는 것임.**

관련없음, 관련있음에 대한 정의는 다양한게 있지만, 그 정의는 변수의 타입에 따라 다름(위 표)

가장 근본적으로 독립성을 검정할 수 있는 것은 Chi-square 기법.

나머지는, 엄밀히 독립성 검정은 아님.

전통적으로는 T-test는 남자 키 평균과 여자 키 평균이 다른지가 궁금해서 데이터를 모으고 쓰는 방법.(두 그룹을 먼저 만들고 키의 차이를 분석함)

빅데이터 시대에는? ⇒ 다른 목적을 위해서 키와 성별을 모으고, 성별에 따라 키를 나눈것

Chi-square

* 두 변수가 카테고리 타입일때!
* O-E/루트(E) ~ N(0,1)에 근사.
* 정규분포에 근사하지 않고 정확히 구하는 것도 있지만(피셔 Exact Test), 컴퓨팅이 엄청나서 2x2 행렬일 때만 사용
* 독립일수록 통계량이 0에 가까워지고
* 독립이 아닐수록 통계량이 커짐

ANOVA - 분산분석(F-test) ⇒ 두 개 이상의 집단에 대한 검정(두 개일땐 T-test도 가능)

* With in group Variance를 각각 구해서 합침
* Between group Variance를 구함
* 그룹내 분산과 그룹간 분산의 차이를 비교하여 그룹간 관련성을 판단