정규분포와 통계적 검정

#표본 #모집단 #통계적검정 #귀무가설 #대립가설 #유의수준
#정규분포 #표준화 #Z-test #T-test

표본과 검정

검정

아침형 인간과 저녁형 인간을 비교하면 유의미한 차이가 있을까? 이러한 질문들에 대해 객관적인 답을 얻기 위해서는 통계를 사용해야 한다. 이 질문을 통계적 방법을 적용할 수 있게 수정하면 이렇게 될 것이다. “아침 7시 전에 일어나는 학생과 그렇지 않은 학생들 간에 학업성취도의 평균이 차이가 나는가?”

통계적 검정은 질문과 추측에 데이터에 기반한 답을 준다는 점에서 의사결정에 매우 중요하다. 단순히 감과 운만으로 결정을 내리기에는 현대의 문제와 선택은 너무나 복잡하다.

표본을 쓰는 이유

어떤 집단에 대한 특성을 알기 위해서 혹은 통계적 검정을 수행하기 위해 집단 전체의 데이터를 사용하는 것은 불가능하거나 매우 큰 비용이 든다. 따라서 통계에서는 연구하려는 집단에서 일정 부분만을 추출해 통계량을 계산하고, 이로부터 확률적으로 원래 집단의 통계량을 예측한다. 이들을 각각 모집단과 표본이라고 한다. 일반적으로 표본의 추출은 주관이나 편향이 개입되지 않도록 난수 등을 이용하는 Random Sampling을 이용한다. 위에서 본 상관관계와 인과관계를 구분하기 위해 임의로 할당하는 것과 같은 맥락이다.

이 때 표본의 통계량, 가령 분산을 계산하기 위해서는 원칙대로는 모집단의 평균을 알고 있어야 하나 이는 불가능하므로, 표본의 평균으로 대체한다. 나아가 분산의 경우 모집단의 분산은 N개의 변량의 편차제곱합을 N으로 나눠 구하는 반면, 표본 분산은 N 대신 N - 1로 나누어 구한다. 이는 자유도라는 개념으로 설명하는데, 표본 분산의 계산에 표본 평균이 쓰이므로, 가질 수 있는 값에 제한이 없는 모집단의 변량과 달리 표본의 경우 값이 변하면서도 같은 평균을 유지해야 하기에 적어도 하나는 평균을 맞추기 위해 고정되므로 N보다 1 작은 값을 자유도로 가지기 때문이다.

통계적 검정의 과정

가설의 설정

통계적 검정은 귀무가설(Null hypothesis)와 대립가설(Alternative hypothesis)의 2가지 가설을 세우는 것에서 시작한다. 예를 들어 동전을 10번 던진 뒤 앞면이 나온 수를 셌더니 7이 나왔다고 하자. 귀무가설은 “동전의 앞면이 나올 확률이 1/2이다”이고, 대립가설은 “동전의 앞면이 나올 확률은 1/2이 아니다”가 된다.

이 때 앞면이 나올 확률이 1/2보다 크거나 작다라는 대립가설을 세우면 단측 검정, 위처럼 어느 쪽인지는 모르겠는데 하여간 1/2과는 다르다는 가설을 세우면 양측 검정이라 하며 이에 따라 p-value를 구하는 방법이 달라진다. 단측 검정은 쉽게 귀무가설을 기각할 수 있지만 연구자의 편향에 의존한다는 문제가 있다.

유의수준 설정

공평한 동전이지만 공교롭게도 10번 다 앞면이 나올 확률도 0은 아니다. 설령 100번을 던져 보아 100번 다 앞면이 나올 확률도 0은 아니다. 통계는 표본의 특성을 이용해 모집단의 특성을 확률적으로 추정하는 것이므로 절대적인 것이 없기에, 확률이 어느 정도여야 가설을 받아들일지에 대한 결정이 필요하다. 이를 유의수준이라 한다.

유의수준은 보통 95%와 99%를 사용하며, 정말 신중한 의사결정의 경우 99%를 쓰고, 일반적으로 95%를 사용한다.

p-value

통계적 검정 기법을 적용해 p-value를 구해낸다. p-value는 귀무가설이 맞다고 했을 때 관찰된 데이터가 나올 수 있는 확률이라고 볼 수 있다. 이 때 p-value가 유의수준에 비추어 작게 나온다면 귀무가설은 기각되고 대립가설이 채택된다. 95%의 경우 p-value < 0.05, 99%의 경우 < 0.01일 때 귀무가설을 기각한다. 즉 대립가설이 해당 유의수준에서 참이라고 판단한다.

정규분포

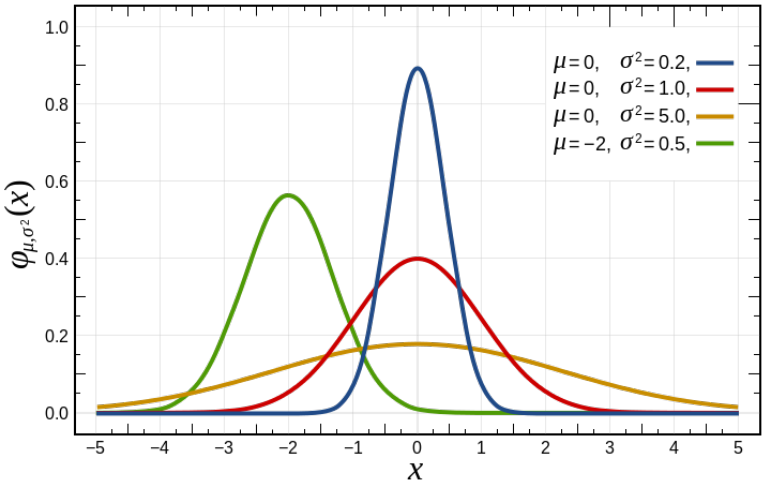
중심극한정리

평균이 μ이고 표준편차가 σ인 모집단에서 크기가 n인 표본을 반복적으로 추출할 때, 이 표본의 평균은 기댓값이 μ이고 표준편차가 $\sigma/\sqrt{n}$ 을 따르는 정규분포에 수렴한다. 또한 확률이 일정하고 독립적인 시행을 n번 반복했을 때 발생하는 사건의 평균 등 확률변수 n개의 평균은 n이 충분히 크다면 정규분포를 따른다는 것이 중심극한정리다.

말은 어렵지만 간단히 생각하면 이전의 결과가 지금의 결과에 영향을 미치지 않는 일을 여러 번 반복하면 해당하는 변수의 분포가 정규분포를 따른다는 것이다. 정규분포는 그 자체의 특이한 성질 외에도, 중심극한정리로 인해 많은 확률변수가 정규분포를 따른다는 점에서 통계학에서 매우 중요하고 자주 쓰인다.

정규분포의 모양과 성질

정규분포의 확률밀도함수는 종의 형태인데, 이 종의 중심은 곧 평균이며, 종의 경사가 완만한지 가파른지는 표준편차에 의해 결정된다.



정규분포와 표준화

평균이 0이고 표준편차가 1인 정규분포를 표준정규분포라 한다. 확률변수 x가 정규분포를 따를 경우, 다음과 같은 식을 사용해 표준정규분포를 따르도록 변형할 수 있다. 이로부터 쉽게 p-value 등을 계산할 수 있다.

$$z = \frac{x - \mu}{s}$$

z-test

위에서 구한 z 값을 이용해 바로 통계적 검정을 할 수 있다. 귀무가설을 "앞면의 확률은 1/2", 대립가설을 "앞면의 확률은 1/2이 아니다", 유의수준을 95%로 설정한 뒤 동전을 100번 던져본 것에 대해 양측 검정을 하자. 귀무가설에 따른 확률분포는 평균 50, 표준편차가 5인 정규분포가 된다. 이 때 앞면이 60번 나왔다고 하면 z값은 (60-50) / 5 = 2가 나온다. 표준정규분포표에서 이 값을 찾으면 0.977250이고 1-0.977250 = 0.02275가 바로 단측검정을 위한 p-value이다. 우리는 지금 양측검정을 하고 있으므로 2배를 해주면 0.0455가 나오며, 95% 유의수준에서 필요한 0.05보다 더 작은 값이 나오므로 귀무가설은 기각된다. 즉, 이 동전은 공평한 동전이 아닌 것이다.

t-test

n이 충분히 큰 경우, 대략적으로 n > 30인 경우 중심극한정리에 따라 정규분포에 가까워지므로 z-test를 쓸 수 있으나, 그렇지 않은 경우 t-test를 사용한다. z-test와 비슷하나 p-value를 구할 때 n의 크기에 따라 다른 값을 사용한다.

One sample t-test = 표본의 평균을 모집단의 평균과 비교하고 싶을 때

Two sample t-test = 두 집단의 평균을 비교하고 싶을 때

Paired sample t-test = 한 집단의 다른 시점의 평균을 비교하고 싶을 때

ANOVA = 비교 집단이 둘 이상인 경우 반복적인 t-test는 오류를 일으키므로 이를 방지하기 위해 ANOVA 사용