빅데이터분석 실무 과제 - 분포

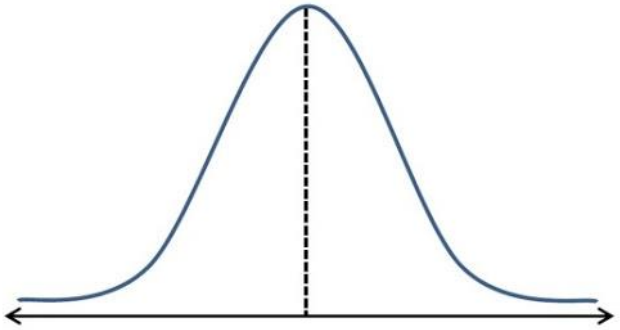
각 분포의 의미

정규 분포

표현: X~N(μ, σ^2)

연속형 변수로서 나타나는 현상을 표현하는 **기본적인 확률 모형**을 의미

ex) 사람의 키, 연봉 등

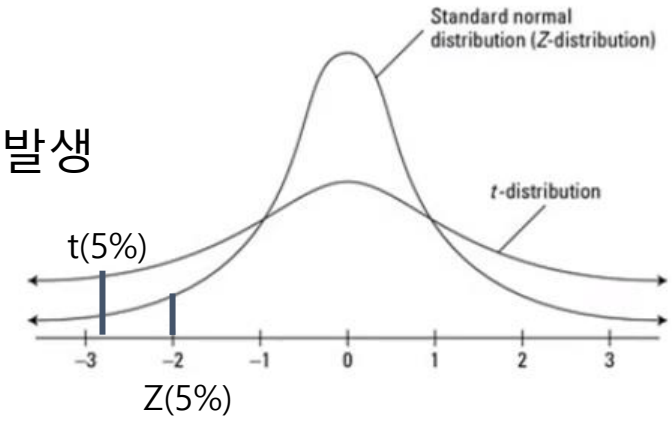


t 분포

표본의 수가 적을 경우 모분산인 σ^2를 알 수가 없음.

-> σ^2가 정확하지 않기 때문에 정규분포로 추정할 경우 **신뢰성**에서 문제가 발생

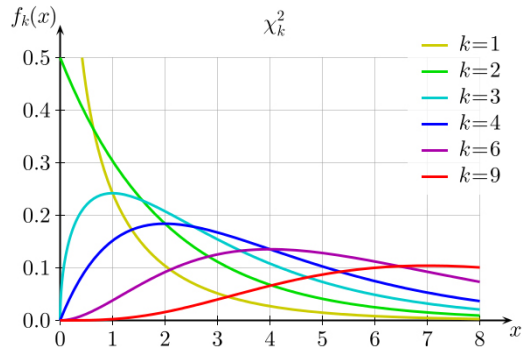
봉우리가 낮고 꼬리가 두꺼운 분포를 사용해서 **웬만한 확신이 없으면 다르다는 결과를 주지 않게 함**.



각 분포의 의미

카이자승 분포

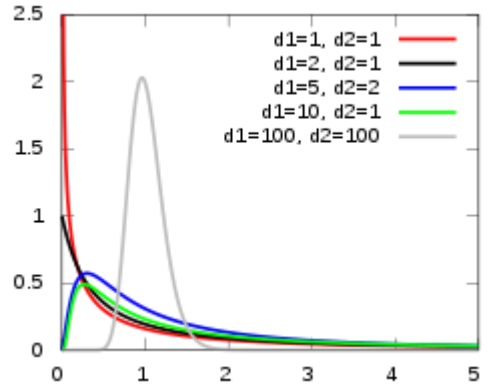
단일 모집단으로부터 추출한 **모집단분산(σ^2)과 표본분산(S^2)의 비**를 나타내는 확률 변수



F 분포

두 모집단에 대한 카이자승 분포 확률변수의 비율로 정의 됨.

$$\frac{\text{첫번째 집단의 표본분산}(S_1^2)}{\text{두번째 집단의 표본분산}(S_2^2)}$$



각 분포 간의 관계

t 분포

$$t = \frac{Z}{\sqrt{\frac{X^2_{(n-1)}}{(n-1)}}}$$

Z → 표준정규분포 확률변수

$X^2_{(n-1)}$ → 카이자승 확률변수

카이자승 분포

$X^2_{(n)} = Z_1^2 + Z_2^2 \ldots + Z_n^2$ → 표준정규분포 확률변수의 합

F 분포

$$F = \frac{\frac{X^2_{(n_1-1)}}{(n_1-1)}}{\frac{X^2_{(n_2-1)}}{(n_2-1)}} \sim F(df_1, df_2)$$

→ 2개의 카이자승 분포의 비율

각 분포의 활용

정규분포

- **평균**에 대한 가설검정에서 많이 사용이 됨.
  가설검정을 시행할 때 정규분포를 이용해서 검정통계량을(Z) 계산

- 회귀분석에서 사용

  잔차의 정규성
  잔차의 분포는 정규분포이어야 한다.

  오차를 분석해 보니 평균이랑 근접할 수록 발생할 확률이 높고, 멀어질수록 확률이 떨어지는 규칙을 보임

  잔차의 분포가 정규분포가 아닌 경우?

  -> 회귀분석이 **설명하지 못하는 특정 패턴**이 있다는 것을 의미

각 분포의 활용

t분포

- **평균**에 대한 가설검정에서 많이 사용이 됨.(**표본수가 적을 때**)
- 회귀분석에서 사용

  회귀계수 검정에서 사용

  $H_0: \beta_0 = 0$: 두 변수 간에는 인과관계(영향력)이 없다.
  $H_1: \beta_1 \neq 0$: 두 변수 간에는 인과관계(영향력)이 있다.

각 분포의 활용

카이자승 분포

- 단일 모집단 **모분산**에 대한 가설검정을 시행할 때 사용이 됨.
  평균에 대해 어느 정도의 산포가 나타나는지를 살펴보는 것
- 교차분석

각 분포의 활용

F분포

- 두 분포의 분산을 비교할 때 활용
- 그룹 내 변동과 그룹 간 변동으로 여러 개의 평균값을 비교할 때 활용(ANOVA)
  분산을 분석하면서 평균을 비교한다.
- 실험계획법
  실험계획법에서 결론을 도출하기 위해서 ANOVA가 쓰인다.
- 회귀분석에서 사용
  회귀모형 자체의 **유의성 검정**을 위해서 사용
  유의성 검정은 회귀직선이 얼마나 의미가 있는가를 보는 것

  $H_0: \beta_0 = \beta_1 = \beta_2 = 0$: 회귀선이 영향력이 없다.
  $H_1: \beta_0 \neq \beta_1 \neq \beta_2 \neq 0$: 회귀선이 영향력이 있다.

  t-test는 **두 집단**의 차이만 분석이 가능하기 때문에 전체 유의성 검정은 F-test로 한다.

참고 자료

"8차시정규분포가그렇게중요한가",

file:///C:/Users/User/Downloads/8%EC%B0%A8%EC%8B%9C%EC%A0%95%EA%B7%9C%EB%B6%84%ED%8F%AC%EA%B0%80%EA%B7%B8%EB%A0%87%EA%B2%8C%EC%A4%91%EC%9A%94%ED%95%9C%EA%B0%80.pdf, 2020-10-17

"t-분포(t-distribution, Student's t-distribution)", https://wikidocs.net/34009, 2020-10-17

"카이제곱분포",
https://m.blog.naver.com/PostView.nhn?blogId=mykepzzang&logNo=220852102307&proxyReferer=https:%2F%2Fwww.google.com%2F

"잔차의 정규성", https://brunch.co.kr/@gimmesilver/17

"회귀분석을 하면 왜 분산분석표(ANOVA)를 보게 되는것일까?",https://m.blog.naver.com/definitice/221333302203

"회귀분석 (Regression analysis)", https://bioinformaticsandme.tistory.com/70

"F-test(F검정)", https://blog.naver.com/vnf3751/220841363022

Thank You