데이터과학과 AI를 위한 파이썬

11강. 추정과 검정

세종사이버대학교

김명배 교수

학습내용

- 분포의 추정
- 가설검정

학습목표

- 주어진 데이터에 대해 이산형 확률분포와 연속형 확률분포들 중 적절한 분포를 추정하고, 분포의 모수를 추정하는 방법론에 대해 설명할 수 있다.
- 귀무가설과 대립가설, 유의수준, 유의확률, 기각역 등의 개념을 설명할 수 있고, 통계적 가설검정에 의해 가설의 채택여부를 판단할 수 있다.

1. 분포의 추정

1) 확률분포를 결정하는 방법

- 데이터 분석 전 첫 번째 가정

  분석할 데이터는 어떤 확률변수로부터 실현된 표본인가?

  → 아직 발생하지 않은 확률변수 값에 대한 발생 확률을 추정할 수 있음

  분포 fitting 또는 분포 모수 추정이라고 함

[분포를 알아내는 방법]

① 확률변수가 어떤 분포를 따를 것인지 탐색을 통해 추정

② 데이터로부터 확률분포의 모수값을 추정

1. 분포의 추정

1) 확률분포를 결정하는 방법

- 데이터 특성에 따른 분포를 추정
  . 변수값이 0 또는 1뿐이다. → 베르누이분포
  . 3개 이상의 범주 값이다. → 카테고리분포
  . 0과 1사이의 실수 값이다. → 베타분포
  . 0 또는 양수이다. → 로그정규분포, 감마분포, F분포, 카이제곱분포, 지수분포, 하프코시분포, 포아송 분포등
  . 데이터가 크기 제한이 없는 실수이다. → 정규분포, 스튜던트 t분포, 코시분포, 라플라스분포

1. 분포의 추정

※ 특수한 분포

모숫값을 조정하여 분포의 모양을 원하는 대로 쉽게 바꿀 수 있어
베이지안 추정에 사용되는 분포

- **베타분포(Beta distribution)**

  표본공간이 0과 1사이로, 베르누이분포의 모수 μ의 값을 베이지안
  추정한 결과를 표현한 분포

- **디리클레분포(Dirichlet distribution)**

  베타분포의 확장판으로, 0과 1사이의 값을 가지는 다변수 확률변수의
  베이지안 모형에 사용됨(베타분포는 k=2인 디리클레분포)

- **감마분포(Gamma distribution)**

  0부터 무한대의 값을 가지는 양숫값을 추정하는데 사용

1. 분포의 추정

※ 특수한 분포

그밖의 연속형 분포

- **로그정규분포(log-normal distribution)**

  데이터에 로그변환한 값이 정규분포가 되는 분포

- **코시분포(Cauchy distribution)**

  스튜던트 t분포에서 자유도(모수)가 1인 경우의 분포

- **하프코시(Half-Cauchy distribution)**

  코시분포에서 양수인 부분만 사용하는 경우의 분포

- **와이블 분포 (Weibull distribution)**

  지수분포를 보다 일반화시켜, 여러 다양한 확률분포 형태를 모두 나타낼 수 있도록 고안된 분포(감마/지수분포의 확장판)

1. 분포의 추정

2) 확률분포의 모수 추정 방법론

- 분포의 모양을 확정하는 모수의 값으로 가장 가능성이 높은 하나의 숫자를 찾아내는 작업을 모수추정(parameter estimation)이라고 함

모수추정 방법론

(1) 모멘트 방법(method of moment)

(2) 최대가능도 추정법(MLE; Maximum Likelihood estimation)

(3) 베이즈 추정법(Bayesian estimaion)

1. 분포의 추정

2) 확률분포의 모수 추정 방법론

(1) 모멘트 방법(method of moment)

- 모멘트가 확률분포의 이론적 모멘트와 같다고 가정하여 모수를 추정

(2) 최대가능도 추정법(MLE; Maximum Likelihood estimation)

- 이론적으로 가장 가능성이 높은 모수를 찾는 방법
- x를 이미 알고 있는 상수로 보고 모수를 변수로 생각함(가능도 함수)
- $L(\theta; x)$로 표시하고 이 가능도를 가장 크게 하는 모수를 찾는 방법

(3) 베이즈 추정법(Bayesian estimaion)

- 모숫값이 가질 수 있는 모든 가능성의 분포를 계산하는 작업
- 모수를 확률변수로 보고 확률밀도함수를 사용함

2. 가설 검정

1) 가설검정이란

- 확률분포에 대한 어떤 주장을 가설(hypothesis)이라고 함
- 이 가설이 어떠한 것이 사실인지를 통계적인 방법으로 증명하는 것을 통계적 가설검정(statistics hypothesis test)이라고 함
- 모집단에 대한 어떤 가설을 설정한 뒤에 통계 기법을 통하여 그 가설의 채택여부를 확률적으로 판정하는 통계적 추론의 한 방법

2. 가설 검정

2) 가설(hypothesis)

- 통계적 가설검정에서 가설은 귀무가설과 대립가설이 있음

(1) 귀무가설(null hypothesis) : H_0

- 기존에 잘 알려지고 증명되어 있는 사실
- 귀무가설은 보통 '~가 같다', '~에 차이가 없다', '0이다' 등으로 표현됨

(예시) 동전의 앞면이 나오는 확률은 0.5이다.

피고인은 죄가 없다(무죄).

성별에 따라 수학 평균점수는 차이는 0이다.

표본의 분포와 정규분포와 차이가 없다(같다).

2. 가설 검정

2) 가설(hypothesis)

(2) 대립가설(alternative hypothesis) : $H_1(not\ H_0)$

- 연구자가 입증하고자 하는 새로운 가설
- 대립가설은 보통 '같지 않다', '다르다', '차이가 있다' 등으로 표현됨

(예시) 동전의 앞면이 나오는 확률은 <u>0.5가 아니다</u>.

피고인은 죄가 <u>있다</u>(유죄).

성별에 따라 수학 평균점수는 <u>0이 아니다</u>.

표본의 분포와 정규분포와 <u>차이가 있다</u>(<u>다르다</u>).

2. 가설 검정

3) 통계적 오류와 유의수준(level of significance)

- 통계적 가설검정은 귀무가설이 옳다는 가정하에 귀무가설이 거짓인 증거가 충분한가를 증명하는 방식임
- 의사결정을 했을 때, 옳은 결정일 수도 있고 잘못된 결정일 수도 있음

가설 선택 \ 진실	귀무가설 진실	대립가설 진실
귀무가설 선택	옳은 결정(신뢰수준(1-α))	**제 2종 오류(β)**
대립가설 선택	**제 1종 오류(유의수준(α))**	옳은 결정(검정력(1-β))

- 제 1종 오류가 더 중요하기 때문에 유의수준(α)을 상수로 고정하여 의사결정을 함($\alpha = 0.05\ or\ 0.1\ or\ 0.01$)
- 즉 유의수준 보다 신뢰성이 있으면 귀무가설을 기각하고 대립가설을 채택함

2. 가설 검정

4) 검정통계량(test statistics)과 유의 확률

- 검정을 위해 특정한 분포를 따르도록 표본으로부터 계산되어 지는 함수의 값을 검정통계량이라고 함
- 검정통계량도 새로운 확률변수의 표본으로 볼 수 있음

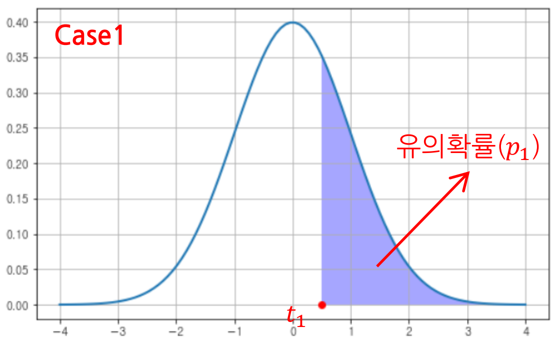
[예시] 정규분포 확률변수에 대한 검정통계량

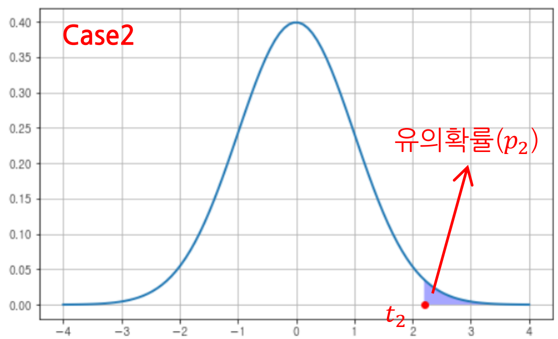
$$x \sim N(\mu, \sigma^2) \quad \rightarrow \quad z = \frac{\bar{\boldsymbol{x}} - \mu}{\sigma/\sqrt{n}} \; \sim \; N(z; 0, 1)$$

2. 가설 검정

3) 검정통계량(test statistics)과 유의 확률

- 표본으로부 터 계산된 검정통계량보다 크게(또는 작게) 나올 확률을 유의확률(p)이라고 함





- 귀무가설 하에 검정통계량 t_1보다 크게 나올 확률(p_1)이 높음

→ 즉, 흔히 일어나는 사건임

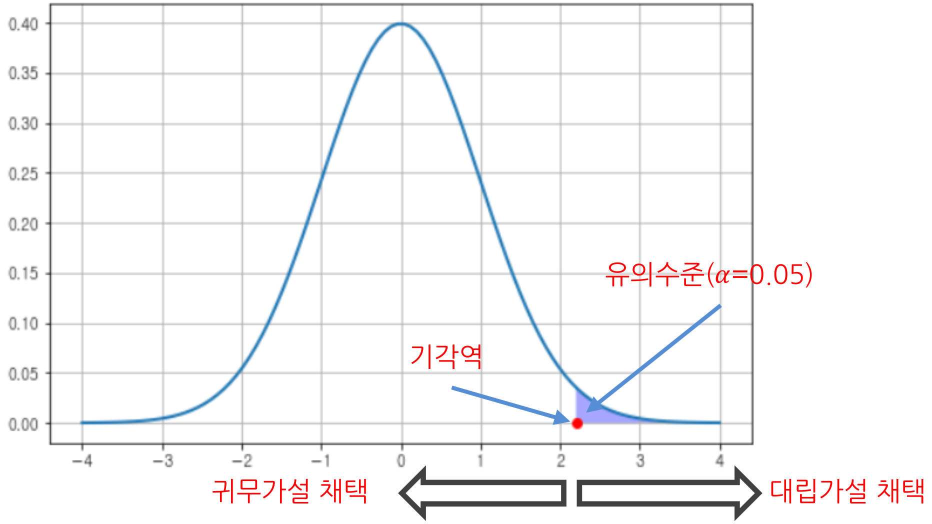
- 귀무가설 하에 검정통계량 t_2보다 크게 나올 확률(p_2)이 낮음

→ 즉, 발생하기 어려운 사건임

2. 가설 검정

4) 기각역과 검정

- 유의수준(α)을 0.05로 하였을 때의 검정통계량 값을 기각역(critical value)이라고 함



2. 가설 검정

4) 기각역과 검정

- 유의수준(α)과 유의확률을 알면 간단하게 가설검정을 할 수 있음

- **유의확률 < 유의수준 ☞ 귀무가설 기각**

- **유의확률 > 유의수준 ☞ 귀무가설 채택**

2. 가설 검정

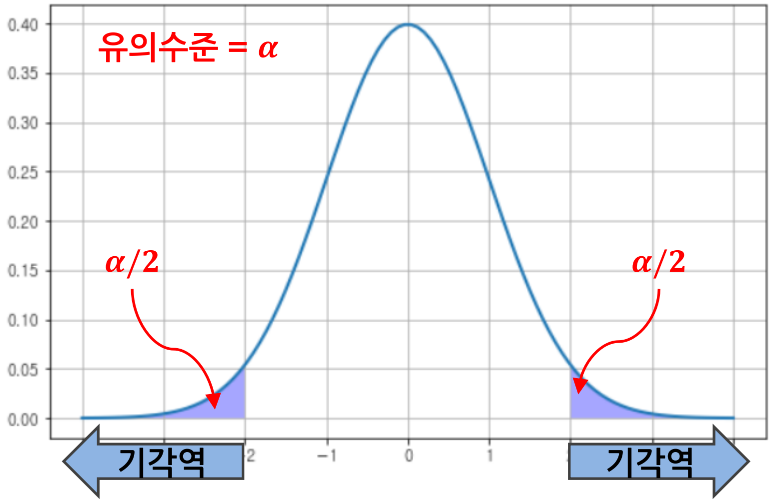
5) 양측검정과 단측검정

(1)양측검정(Two-side Test)

검정통계량의 분포에서 기각영역이 양쪽에 나타나는 형태의 가설검정

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$



2. 가설 검정

5) 양측검정과 단측검정

(2)단측검정(one-side Test)

검정통계량의 분포에서 기각영역이 양쪽에 나타나는 형태의 가설검정

- 좌측검정(left-side test)

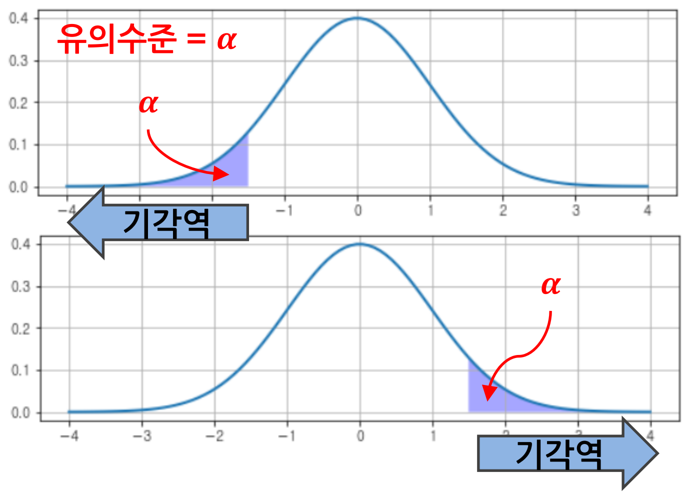
$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0$$

- 우측검정(right-side test)

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$



2. 가설 검정

5) 양측검정과 단측검정

[예시] 교사에 대한 태도 검사를 실시해 온 결과, 작년까지 평균은 178이었고 표준편차는 24이었다. 금년에도 36명의 임의표본을 추출하여 동일한 검사를 실시한 결과 평균이 170이었다. 이 결과로부터 학생들의 교사에 대한 태도는 변하였다고 할 수 있는지 검증하세요.

[풀이]

- 가설 $H_0 : \mu = 178,\ H_1 : \mu \neq 178$
- 유의수준 $\alpha = 0.05$
- 검정통계량 : $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
- H_0가 참이라고 할 때 검정통계량의 표집 분포 $z \sim N(0,1)$
- 기각역 $z > 1.96$ 또는 $z < 1.96$
- $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{170-178}{24/\sqrt{36}} = -\frac{8}{4} = -2.0$

➔ 검정통계량 값이 기각역에 들어가기 때문에 H_0기각, H_1 채택

2. 가설 검정

5) 양측검정과 단측검정

[예시] 은행이 한 시간 동안 평균 10명의 고객을 응대하고 있고, 표준편차는 2.3이다. 시스템 도입 이후 총 50명의 직원들을 대상으로 고객 응대의 시간을 조사해본 결과 한 시간 평균 10.4명의 고객을 응대하고 있다. 고객응대시스템 도입을 통해 한 시간 동안 처리하는 고객의 수가 늘었다고 할 수 있는지 검증하세요.

[풀이]

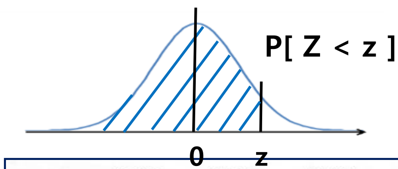
- 가설 $H_0 : \mu = 10 \; H_1 : \mu > 10,$ 유의수준 $\alpha = 0.05$
- 검정통계량 : $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
- H_0가 참이라고 할 때 검정통계량의 표집 분포 $z \sim N(0,1)$
- 기각역 $z > 1.645$
- $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{10.4-10}{2.3/\sqrt{50}} = 1.23$

➔ 검정통계량 값이 기각역 보다 작기 때문에 H_0채택, H_1 기각

2. 가설 검정

5) 양측검정과 단측검정

[예시] 표준정규분포표(일부)



단측 검정 : (1-유의수준) 에 해당되는 z값 찾기
양측 검정 : (1-유의수준/2)에 해당되는 z값 찾기

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936

정리하기

1. 가설검정

- 분포를 알아내는 방법
  ① 확률변수가 어떤 분포를 따를 것인지 탐색을 통해 추정
  ② 데이터로부터 확률분포의 모수값을 추정
- 데이터 특성에 따른 분포를 추정
  - 변수값이 0 또는 1뿐이다. → 베르누이분포
  - 3개 이상의 범주 값이다. → 카테고리분포
  - 0과 1사이의 실수 값이다. → 베타분포
  - 0 또는 양수이다. → 로그정규분포, 감마분포, F분포, 카이제곱분포, 지수분포, 하프코시분포, 포아송 분포등
  - 데이터가 크기 제한이 없는 실수이다. → 정규분포, 스튜던트 t분포, 코시분포, 라플라스분포
- 모수추정 방법론

(1) 모멘트 방법(method of moment)

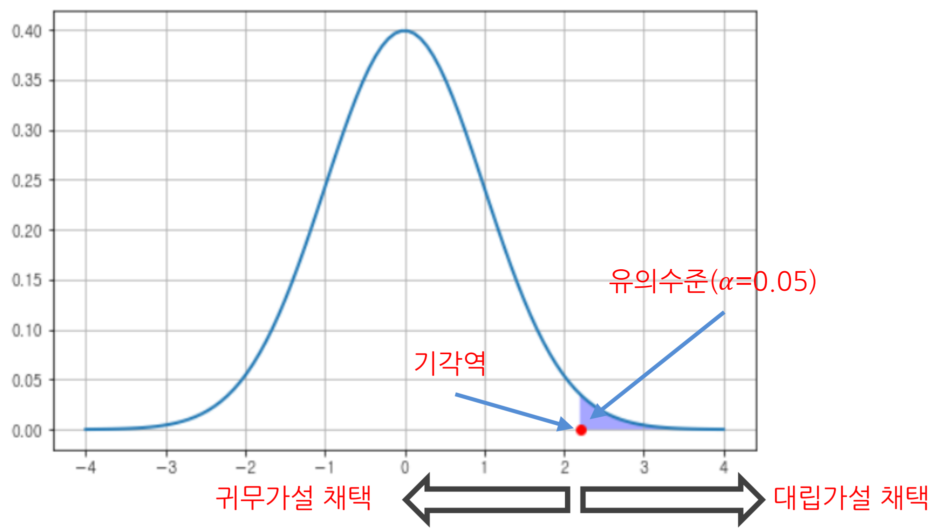
(2) 최대가능도 추정법(MLE; Maximum Likelihood estimation)

(3) 베이즈 추정법(Bayesian estimaion)

정리하기

2. 가설검정

- 가설이 어떠한 것이 사실인지를 통계적인 방법으로 증명하는 것을 <u>통계적 가설검정(statistics hypothesis test)</u>이라고 함
- 가설 : 귀무가설과 대립가설이 있음
- 유의수준 : 제 1종류로 귀무가설이 진실인데 기각할 오류
- 유의확률 : 검정통계량보다 더 크거나 작게(보다 희귀하게) 나올 확률



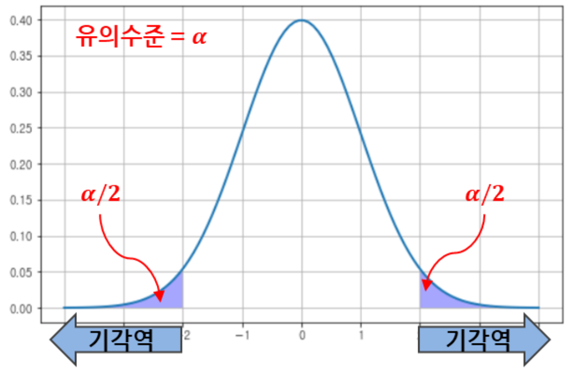
정리하기

2. 가설검정

- 양측검정 : 검정통계량의 분포에서 기각영역이 양쪽에 나타나는 형태의 가설검정

$H_0: \mu = \mu_0$

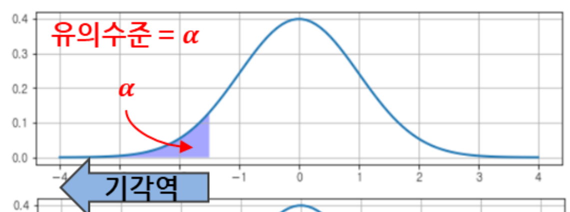
$H_1 : \mu \neq \mu_0$



- 단측검정 : 검정통계량의 분포에서 기각영역이 양쪽에 나타나는 형태의 가설검정
  - 좌측검정(left-side test)

    $H_0: \mu = \mu_0$

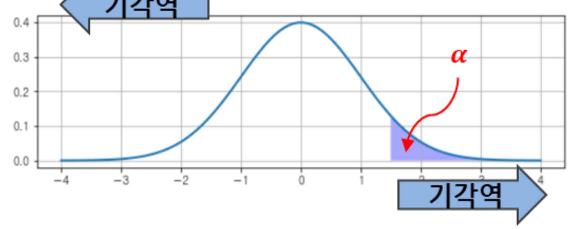
    $H_1 : \mu < \mu_0$



  - 우측검정(right-side test)

    $H_0: \mu = \mu_0$

    $H_1 : \mu > \mu_0$



정리하기

2. 가설검정

1. 귀무가설과 대립가설의 수립	▪ 귀무가설 : 기존의 사실, 기존에 받아들이던 가설 ▪ 대립가설 : 표본을 통해 새롭게 입증하고자 하는 가설
2. 유의수준 설정	▪ 유의수준 : 제 1종 오류(귀무가설이 참인데, 대립가설을 선택하는 오류), 보통 5% 기준으로 사용한다.
3. 통계적 분석 기법의 선택	▪ 독립변수와 종속변수의 척도(범주형 or 연속형)에 따라 확률분포를 적절하게 선택한다.
4. 검정통계량 vs 기각역 유의확률 vs 유의수준	▪ 검정통계량과 기각역 또는 유의확률과 유의수준의 대소관계를 판단한다.
5. 귀무가설 기각 여부 결정	▪ 유의확률이 유의수준(사용자 결정)보다 작거나, ▪ 검정통계량이 기각역보다 크면 귀무가설을 기각한다.
6. 최종 결론 및 의사결정	▪ H_0또는 H_1 기각 여부를 판단하여 최종 의사결정을 한다.