집고 넘어가기!

추정

표본을 이용하여 모집단의 미지의 값을 추측하는 과정

Ex) 우리나라 대학생 평균키는 173이다.

Ex) 20대 취업자 연봉은 3000이다.

가설검정

귀무가설 : 기존에 알려져 있는 사실 (전재로 하는 가설)

대립가설 : 새로운 사실, 뚜렷한 증거로 입증하려는 가설

모집단의 주장 또는 가설의 옳고 그름을 판단하는 것

Ex) 귀무가설 : 구슬아이스크림의 용량은 50ml가 맞다

대립가설 : 구슬아이스크림의 용량은 50ml가 아니다

이러한 가설검정의 과정중 통계에서는 귀무가설만 검정한다.(알파)

귀무가설은 알려진 사실이기 때문에 명확히 가설로 진술할 수 있기 때문이다

신뢰구간은 유의수준과 같은 의미이다.

알파(a) 는 1종오류 (귀무가설이 참인데 대립가설을 채택하는 오류)

귀무가설이 진실인데 기각할 수 있는 오류를 범할 확률의 최대 허용치 (0.05)

쉽게 요약하면 대립가설을 채택할 때 범할 오류가 5%라는 의미 (5%보다 작으면 대립가설 선택)

해석 : 귀무가설을 기각하고 대립가설 채택 확률이 5% 미만이다

문제로 확인하기

[2015 학년도 수능] 어느 연구소에서 토마토 모종을 심은 지 주가 지났을 때, 줄기의 길이를 조사한 결과 토마토 줄기의 길이는 평균 30cm 표준편차가 2cm 인 정규분포를 따른다고 한다. 이 연구소에서 토마토 모종을 심은 지 주가 지났을 때, 토마토 줄기 중 임의로 선택한 줄기의 길이가 27cm 이상이고 32cm 이하일 확률을 오른쪽 표준정규분포표를 이용하여 구한 것은? [3 점]

1) 문제에서 z에 관한 표가 나왔습니다. 그러므로 이 문제는 정규분포 관련 문제입니다.

2) 문제에서 평균과 분산을 찾아 $N(m, \sigma^2)$ 형태로 적습니다. 예제에서는 평균이 30, 표준편차가 2이므로 $N(30, 2^2)$라고 적으면 됩니다.

3) "임의로 선택한 줄기의 길이가 $27cm$ 이상이고 $32cm$ 이하일 확률"을 식으로 적으면 $P(27 \leq X \leq 32)$ 입니다.

$$P(27 \leq X \leq 32) = P(\frac{27-30}{2} \leq Z \leq \frac{32-30}{2})$$

$$= P(-1.5 \leq Z \leq 1) = 0.4332 + 0.3413 = 0.7745$$

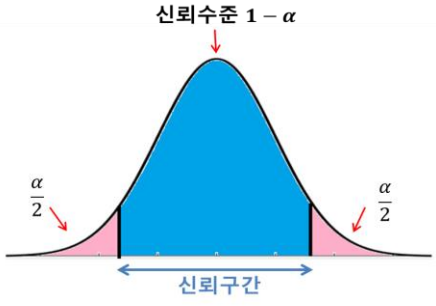
로 답이 나오게 됩니다.

여기서 문제를 간단하게 변형! 가설 과정으로 변경! 임의로 선택을 한 줄기가 27cm

가설 설정

귀무가설 : 토마토 줄기의 길이는 평균 30CM 이다

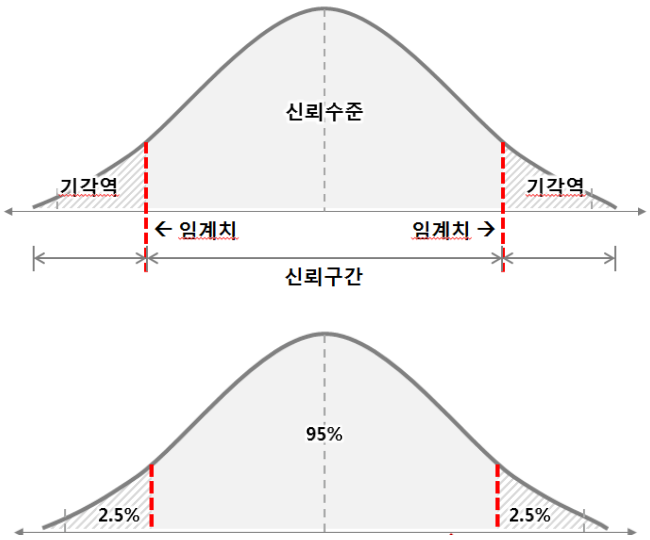
대립가설 : 토마토 줄기의 길이는 평균 30CM 가 아니다



위의 가설을 검정한다고 가정을 한다면 이 검정은 양측검정이며 30 보다 클수도 있고 작을수도 있다는 경우의 수를 열어두고 검정하는 가설이다 그렇기에 양쪽의 a(알파)/2 를 통해서 양쪽의 0.025 씩 나눠갖는다고 생각하면 된다

그렇게 되면 1-a 는 95%의 신뢰구간이 만들어지며 유의수준 또는 임계치의 값을 알 수 있다

추가적인 그림



표준정규본포 표에 의하면 0.025 의확률은 같는 임계치는 1.96 이다. (별첨. 표준정규분포포 참고)

그렇기에 평균이 30 이라고 말할 수 있는 95%신뢰구간을 구해보면

(X-30) / 2 = 1.96 이 되는 값이므로 X 는 33.92 가 나오며 30 에서 3.92 라는 값의 차이가 나오게 된다. 즉 30 + 3.92 = 33.92 와 30-3.92 = 26.08 이 두가지 값이 신뢰구간 95%안의 값 (임계치) 이다

그렇다면 임의로 표본을 뽑았을 때 값이 26 이라고 가정을 한다면 이건 유의수준(신뢰구간) 0.05 보다 작으므로 (95%신뢰구간 범위를 벗어남으로) 대립가설을 채택

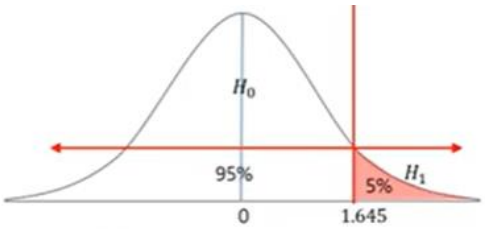
즉, 토마토 줄기의 평균은 30 이라고 말할 수 없다. 라는 결론이 도출된다!

(물론 표본을 1 개를 뽑아서 하는 것이 아니라 여러 개를 뽑았을 경우는 표준편차를 표본의 크기 루트값으로 나눠주게된다)

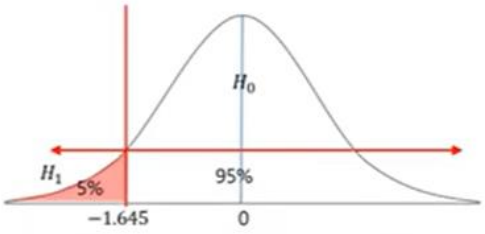
여기서 중요한 것은 검정의 연산식이 어떻게 되는 것이 아니라! 가설 검정에 대해서 귀무가설을 채택하는지 기각하는지에 대한 기준을 보는 방법을 익혀야된다!

양측검정과 단측검정

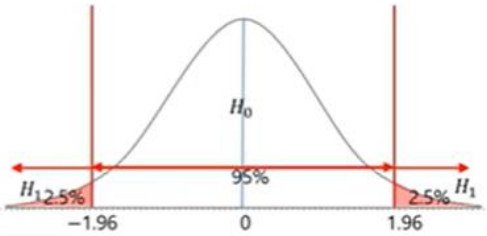
- α=0.05일 때
  - 우측검정(right-sided test)



  - 좌측검정(left-sided test)



  - 양측검정(two-sided test)



우측

귀무가설 : 토마토 줄기의 길이는 평균 30CM 이다

대립가설 : 토마토 줄기의 길이는 평균 30CM 보다 크다

좌측

귀무가설 : 토마토 줄기의 길이는 평균 30CM 이다

대립가설 : 토마토 줄기의 길이는 평균 30CM 보다 작다

이와 같이 단측검정을 하게된다면 그 반대쪽은 검정을 하지 않아도 된다. 따라서 95%신뢰구간을 구할 때 양쪽에 0.025 씩을 나눠갖는게 아니라 0.5 를 갖게되는 형태이다.

이때 위에서 양측 95%신뢰구간 값은 1.96 으로 했지만 단측인경우는 1.645 값을 하면 된다

정리를 해보면!

아이스크림 50ml용량이 맞는지를 검정을 한다.

가설

귀무가설 : 아이스크림 용량은 50ml이다

대립가설 : 아이스크림 용량은 50ml가 아니다

통계검정값의 신뢰구간은 (45~55)이라고 가정한다

-위에서도 얘기했듯이 검정자체에 대한 식보다 과정을 이해하는게 중요하다

-R에서는 유의확률을 바로 구해주거나 신뢰구간을 바로 구할 수 있다.

그럼 실제 표본을 뽑았을 때 44가 나왔으면 신뢰구간에서 벗어나게 된다

(유의확률이 0.05보다 낮다라고 이해하면 된다!, 신뢰구간이 유의확률 0.05보다 낮을때의 최대 허용치이기 때문이다)

따라서 대립가설 채택

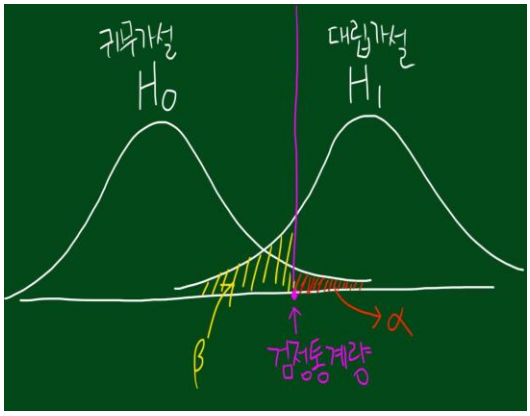
즉, 표본을 뽑았을때 값이 44로 유의확률이 유의수준 0.05보다 낮으므로 대립가설을 채택한다

따라서 아이스크림 용량은 50ml가 아니다

(유의확률을 구하는 방법은 손으로 구하는 것이 아니라 표준정구분포표 등을 보면서 구하는 것이며 추후에 해석하거나 의사결정할때는 유의확률을 보고 결정을한다)

(위에 범위문제가 주어졌을경우처럼 유의확률은 77.45인 형태처럼 95% 신뢰구간의 범위를 구했던 것이다!)

추가!



어떠한 가설 검증이던 판단착오의 위험이 무조건 따른다!

H0은 아이스크림용량은 50이라는 분포이고

H1은 아이스크림 용량이 56이라는 값이 나왔을때의 분포라고 생각하면 된다!

그럼 여기서 H0를 사실이라는 가정하에서 검정하기 때문에 (a)알파를 보고

알파는 일종오류의 확률이라고 생각하면 된다!

즉, 여태까지 얘기했던 0.05보다 낮다라는건 1종오류 확률이 0.05보다 낮으니까 대립가설이 맞다라고 하는 것이다

근데 여기서b(베타) 라는것도 존재하며 대립가설이 만약 참이라고 한다면 귀무가설을 택했을때의 오류 확률을 계산하는 식이다. 통계검정을 할때는 잘 사용되지 않는다

아지만 PPT에서 나왔다시피

1종오류가 무엇이고 2종오류가 무엇인지 정도만 이해하며 의미를 알고있는 것이 중요!

일종오류 : 귀무가설이 참이였을때 대립가설을 선택해서 오류를 범할 확률

이종오류 : 대립가설이 참이였을때 귀무가설을 선택해서 오류를 범할 확률

	귀무가설 채택	귀무가설 기각
귀무가설 참	옳은 결론	잘못된 결론 (알파)
귀무가설 거짓	잘못된 결론 (베타)	옳은 결론

가설검정 참고링크

https://kkokkilkon.tistory.com/36

이 링크는 위에서 제시된 Z검정이 아닌 T검정으로 진행한것이므로 검정자체에 대한부분모다 가설검정의 과정을 이해!

중심극한정리 참고링크

https://angeloyeo.github.io/2020/09/15/CLT_meaning.html

중심극한정리는 필수로 암기해야되는건 아니지만 알고있으면 매우 유용함!

마지막! R함수 해석

6.1 난수함수(rnorm)

난수함수는 정규분포함수의 변수에 해당하는 값을 임의로 생성해 주는 함수

```
> rnorm(5)
[1] -1.1819541 -0.6065962  0.6924985 -0.8988901  0.5788439
```

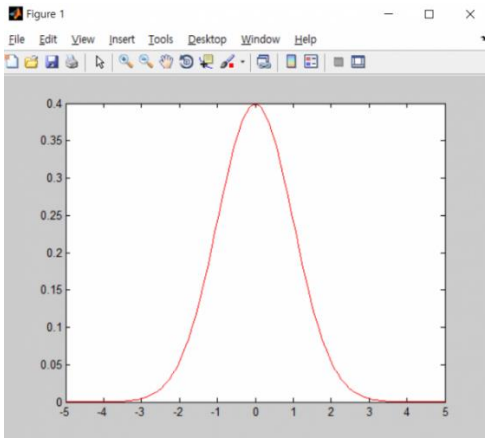
평균을 100, 표준편차를 5로 바꿔봅시다.

```
> rnorm(5,mean=100,sd=5)
[1] 104.68508 106.81284  96.29777 101.26942 109.49112
```

6.2 확률밀도함수(dnorm) -> **불필요!**

확률밀도함수의 함수값을 구해줍니다. 확률밀도함수이기 때문에 값 자체가 확률을 의미하지는 않습니다. 디폴트 평균이 0이니까. 최댓값은 0에서 발생합니다.(확률정규분포이해용!)

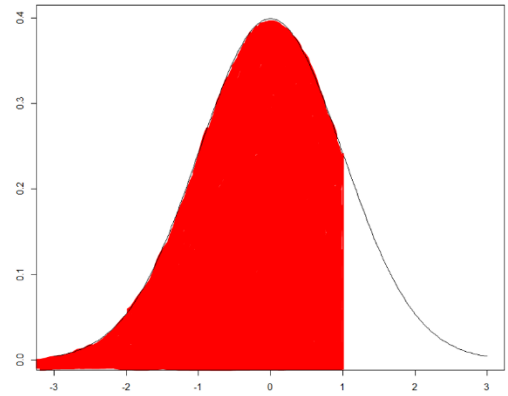
```
> dnorm(0)
[1] 0.3989423
```



x축에는 확률변수의 값을 y축에는 확률변수가 갖는 구간값이 나타날 확률을 표현한 곡선이고 x값에 따른 y의 나타날 확률의 값을 표현! 이론 상으로는 연속적으로 정의되지만, 실질적으로는 실험적으로 얻어진 한정된 샘플에 의해서 정의되며 전체 샘플 수에서 이산화된 구간 내의 사건이 발견될 확률을 히스토그램(histogram)으로 표현된다

6.3 누적분포함수(pnorm)

여기서 나오는 값은 실제 우리가 구한 검정통계량값이 어느정도의 유의확률이 나오는지를 구할 수 있다!

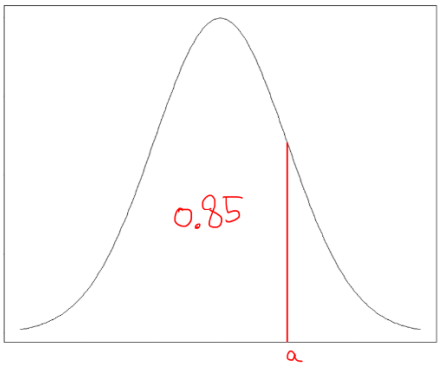


```
> pnorm(1)
[1] 0.8413447
```

실제 1.96을 넣으면 0.975라는 값이 나오게 되는것이다! 밑에 qnorm과 반대의 개념

6.4 분위수함수(qnorm)

확률이 입력변수이고, 어떤 확률을 입력하면 그 확률에 해당하는 변수를 찾아준다. (위에서 나왔던 1.96, 1.645가 여기서 나오는 값)




```
> qnorm(0.85)
[1] 1.036433
```

0.95를 치면 1.645가 나올것이다! 0.975를 치면 1.96이 나오며 양층검정 단측검정에 대한 이해를 했다면 이제 명확이 이해할 것이다!