통계학 입문

2021-05-16

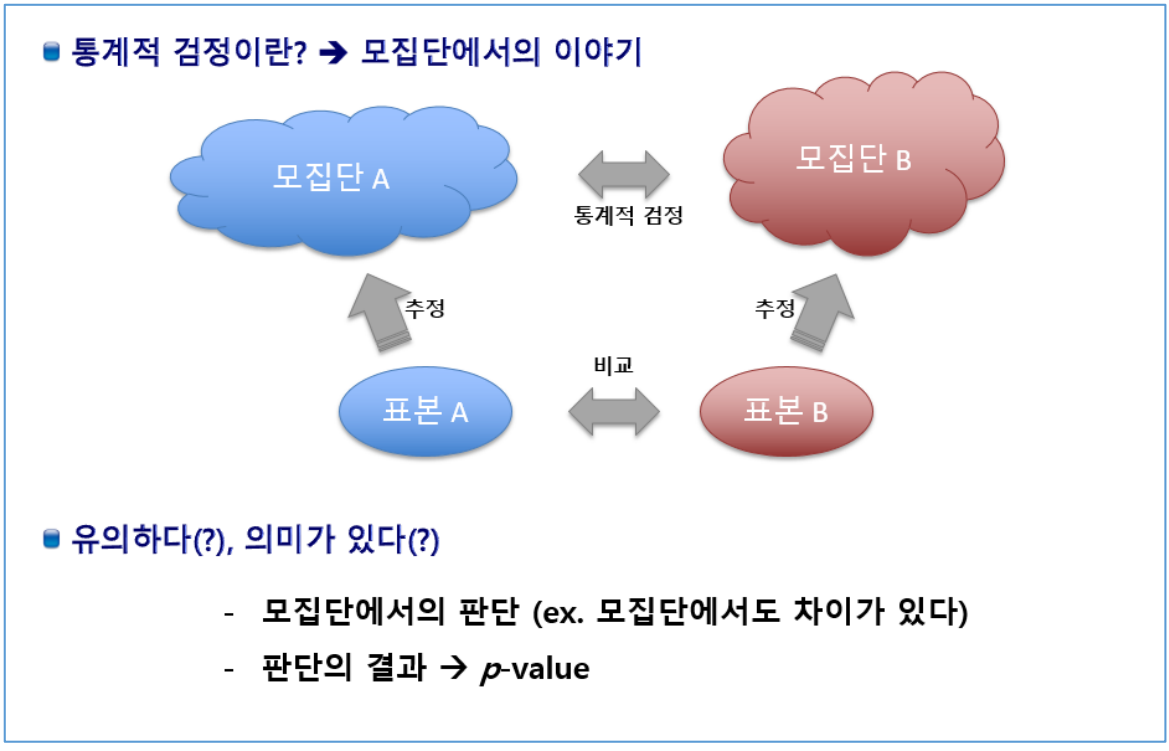
데이터과학융합스쿨

손 대 순 교수

지난 시간에 강의한 '통계의 꽃'에 대해 어렵다는 원성이 자자하네요. 그래서, 오늘은 이 부분에 대해 글로 다시 한 번 이야기 하려고 합니다. 영상으로 보는 것과 글로 이해하는 것은 다른 느낌일테니, 천천히 다시 한 번 생각해 보도록 합시다.

통계적 검정 과정에는 매우 중요한 용어들이 있어요. 아마 여러분들의 전공분야를 공부하면서도 많이 등장하는 용어일겁니다. 정확하게 개념을 이해하는 것이 중요해요. 검정의 기본에 대한 간단한 실습도 해 보려고 합니다. 지난주와 오늘의 내용을 명확하게 알고 있어도 검정의 기본은 알고 있다고 할 수 있어요.

검정(test)이 무엇일까요?



검정은 표본에서 관찰되는 현상이 모집단에서도 존재한다고 할 수 있을지에 대한 과학적 추론 과정입니다. 표본에서 두 집단의 평균 차이가 1만큼 난다고 해 보죠. 과연 모집단에서 차이가 있

다고 할 수 있을까? 이에 대해서 합리적으로 생각해 보는 과정을 의미합니다.

예를 한 번 들어봅시다. 2000년도에 우리나라 인구의 평균 신장이 160cm라는 통계를 가지고 있다고 합시다. 현재 시점에서 인구 1000명을 표본을 뽑아 키를 측정하였더니, 평균이 165cm가 계산되었습니다. 과거에 비해 평균 키가 커진 것일까요? 이에 대한 검정을 해 보겠습니다.

가정 먼저 해야 할 일은 가설을 세우는 일입니다. 위의 경우는 다음과 같은 가설을 세우게 됩니다.

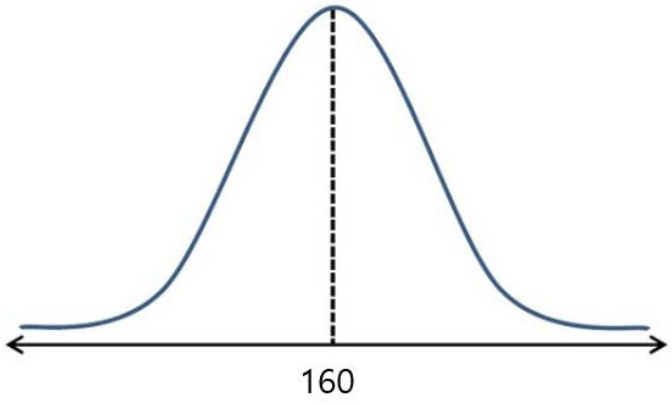
귀무가설 (H_0) : 현재 우리나라 국민의 평균 키는 160cm 이다. ($H_0: \mu = 160$)

대립가설 (H_1) : 현재 우리나라 국민의 평균 키는 160cm 보다 크다. ($H_1: \mu > 160$)

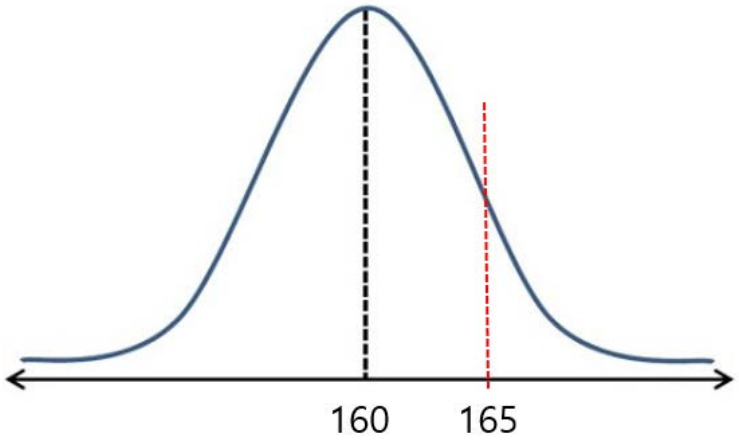
귀무가설은 영가설이라고도 하고, 영어로는 null hypothesis라고 합니다. 기호로는 H_0를 사용하죠. 이와 반대되는 가설을 대립가설, alternative hypothesis라고 하고, 기호로는 H_1이나 H_a를 사용합니다. 일반적으로 연구자가 주장하고자 하는 가설을 대립가설에 둡니다. 위의 예에서 보면, 표본의 평균이 165가 나왔으니, 알고 있는 모집단의 평균, 160보다 크다는 심증을 갖게 되죠. 이를 판단하고자 검정 과정을 진행하는 것이니, "160보다 크다"를 대립가설로 두고, "모집단의 평균이 알고 있는 160 그대로다" 하는 것을 귀무가설로 두는 것이죠.

모든 검정은 "귀무가설이 맞다"는 것을 가정하고 있습니다. 귀무가설이 참이라 할 때 현재 손에 쥐고 있는 데이터(표본)와 같은 상황이 발생할 확률을 계산하는 것이죠. **이 확률이 크다**면, '귀무가설이 참이라고 할 때, 이런 데이터 상황이 발생할 가능성이 충분히 높으므로 굳이 귀무가설이 틀렸다고 할 수 없다'라는 결론을 내리게 되는 것이죠. 만일 **이 확률이 매우 작다**면 '귀무가설 하에서 이러한 일은 발생할 가능성이 매우 낮으므로 이는 귀무가설이 잘못되었다고 판단'하는 것이 타당하다고 보는 것이죠. 이것이 검정의 프로세스예요. 따라서 그 첫 단계인 가설의 설정은 매우 중요합니다. 모든 통계적 검정은 이러한 가설을 가져야 할 수 있습니다. 그래서 '가설 검정'이라고도 합니다.

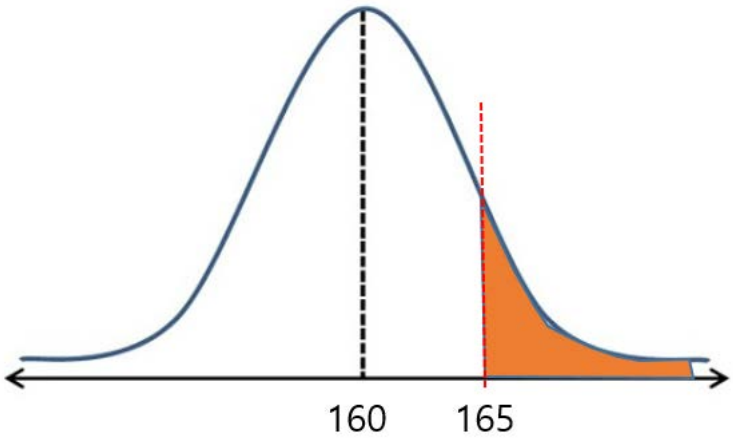
가설 검정에서 다음 스텝으로 할 일은 확률을 계산하는 일이겠지요. 모집단의 분포와 상관없이,



n이 크다면 표본평균은 위와 같은 분포를 가지겠지요 (중심극한정리). 귀무가설을 참이라고 생각하기로 했으니 위와 같이 평균 160을 중심으로 한 분포일 것입니다. 실제 표본의 평균이 165가 계산되었다네요. 165의 위치를 이 분포에 표시해 보면 다음과 같습니다.

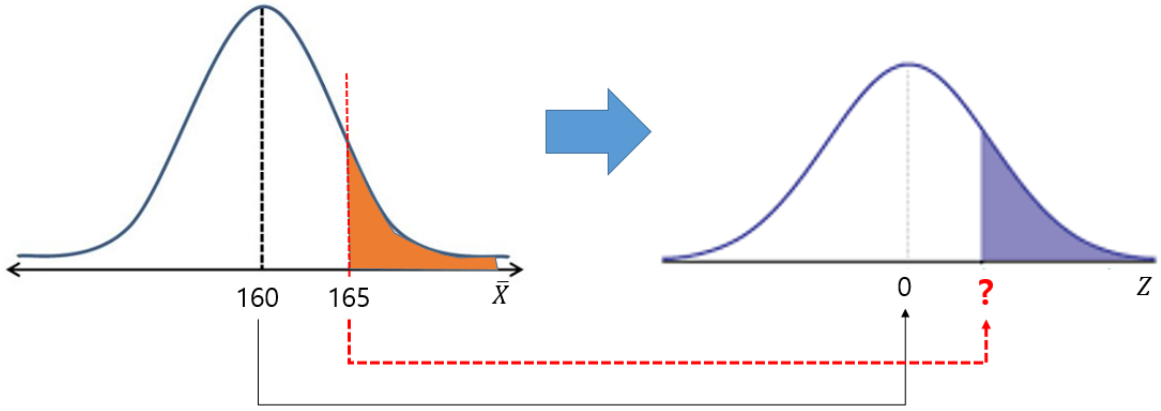


평균이 160인 분포에서 165인 경우의 위치를 표시한 것이예요. 그렇다면 165보다 큰 부분의 면적이 어떤 의미를 가지고 있는지 생각해 봅시다.



위의 색칠한 면적이 무엇을 의미할 까요? 이는 모평균이 160인 분포에서 표본을 뽑았을 때, 표본평균이 165보다 클 확률을 의미합니다. 귀무가설이 참일 때를 가정한 분포에서 현재 데이터 상황 (표본 평균이 165인 상황) 보다 더 극심한 상황 (표본 평균이 165보다 큰 상황)이 발생할 확률이 되는 것이죠. 이것을 p-value (p값)이라고 합니다. 여러분들이 앞으로 귀가 따갑게 들을 용어예요. p-value의 정의에 대해서는 여러가지 방향으로 설명할 것입니다. 그 중 하나가 위의 설명입니다.

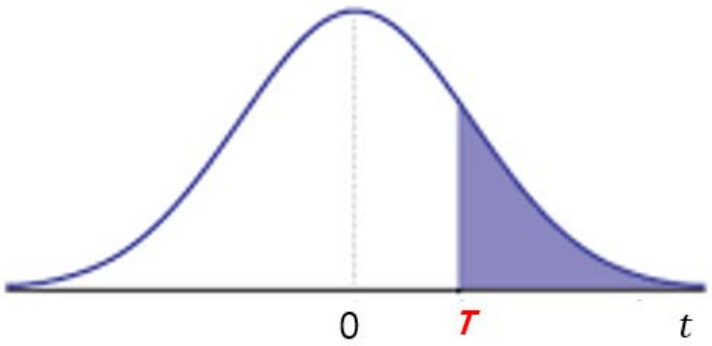
그렇다면 이 p-value를 어떻게 계산할 수 있을까요? 그렇죠. 표준정규분포로 정규화하여 확률을 계산할 수 있습니다.



표준정규분포라는 것은 평균이 0이고 분산이 1이죠. 그러니 표준화를 시키면 평균 160은 0이 될 것입니다. 165는 어떤 값이 될까요? 이 값만 정확히 알면, 표준정규분포표를 이용해 그 값보다 큰 부분의 면적을 계산할 수 있을 것이고, p-value도 구할 수 있겠죠. 이를 표준화 하는 일은 정규분포 단원에서 열심히 연습했던 그 식입니다.

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

이 식이 이제 눈에 잘 들어오나요? 모르는 기호는 없죠? 표본 1000명의 평균이 165라고 했으니 $\bar{X}=165$ 입니다. 귀무가설이 참이라는 가정을 하고 분석한다고 했죠? 따라서, $\mu=160$이예요. 귀무가설을 그렇게 설정했으니까요. $n=1000$도 알고 있는 값입니다. 앞에서 잠깐 언급한 대로 모표준편차, σ를 알고 있으면 이를 채워 넣고 표준정규분포표를 이용해 확률을 계산하면 되고, σ를 모른다면 표본표준편차를 이용하고, t-분포표를 찾으면 된다고 했죠. 대부분의 경우는 모분산을 알지 못하니, t-분포를 많이 이용합니다. 정리해 보면

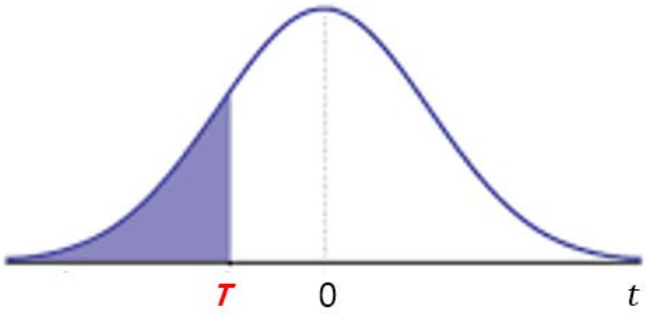


결국 T값을 알면 이를 기준으로 확률계산이 가능하다는 이야기죠. T값은 다음과 같이 계산합니다.

$$T=\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

이 것이 one-sample t-test입니다. 위 식의 T를 one-sample t-test의 검정통계량 (test statistic)이

라고 해요. 결국 이 검정통계량이 크면, p-value는 작아지는 것이죠. 조금 엄밀하게 말하면 아래의 그림과 같이 "OOO보다 작다"를 검정하는 경우도 있습니다. 예를 들어 표본의 평균이 150이 나왔다면, 대립가설이 $H_1: \mu < 160$로 세워지겠지요. 그렇다며 T값은 0보다 작은 값이 되고, p-value도 아래 그림에서의 면적으로 정의됩니다.



그러니, 검정통계량이 커지면 늘 p-value가 작아진다고 할 수는 없습니다. p-value는 양 극단의 확률이므로, 검정통계량의 절대값이 커지면 작아진다고 해야 맞는 표현이겠지요.

자, 이제 판단이 남았습니다. 이 정도면 귀무가설처럼 모집단의 평균은 160라고 할 수 있을까요? 아니면 대립가설처럼 모집단의 평균이 160보다 크다고 해야 할까요? 어떻게 판단하는 것이 좋을지 생각해 보고자 합니다. 앞에서 언급한대로, p-value가 작으면, 귀무가설하에서 이러한 일이 발생할 확률이 작은 것이니, 귀무가설이 잘못되었다고 판단하는 것이 타당하겠지요. p-value가 크다면 귀무가설이 잘못되지 않았다는 말일 테니 아쉽지만 귀무가설을 받아들여야겠지요. 여기에서 중요한 단어는 크다/작다입니다. 어느 정도가 작은 것이고, 어느 정도가 큰 것인지 기준이 있어야 판단을 하겠지요. 그 기준을 우리는 '유의수준(Significance Level)'이라고 합니다. 우리가 정하는 유의수준이라는 기준을 두고 이보다 작으면 대립가설을 취하게 되는 것이죠. 일반적으로 유의수준은 5%를 많이 사용합니다. 즉, 5%보다 p-value가 작으면 'Reject Ho', H0를 기각한다는 표현을 하고, 이보다 크면 'Do not reject H0', 귀무가설을 기각하지 못한다라는 표현을 쓰는 것이죠.

이는 마치 판사가 피고인을 판결하는 것과 유사합니다. 기본적으로 피고인이 무죄라는 가정을 하고 재판을 시작하죠. 이는 귀무가설이 참이라는 가정과 유사합니다. 검사는 피고인이 유죄임을 입증하기 위해서 애를 씁니다. 판사는 그 증거가 충분하다고 생각한다면 피고인을 유죄로 판단하게 되는 것이죠. 죄가 없다는 귀무가설을 기각하는 것입니다. 이 정도면 유죄판결을 내리기에 충분하다고 생각하는 판사의 기준이 바로 유의수준에 해당하는 것이죠.

'유의하다'라는 말을 많이 사용합니다. '표본에서의 차이가 유의한가'라는 질문으로 '유의하다' 혹은 '유의하지 않다'라고 답을 합니다. 이 것의 의미는 '모집단에서도 차이가 있다고 할 수 있는가?'라는 의미입니다.

모든 통계적 가설검정은 이와 같은 방식으로 진행합니다. 검정의 결과는 p-value로 나오는 것이죠. 이 p-value와 유의수준을 비교하여 유의수준보다 작으면 미리 설정해 놓은 귀무가설을 기각하는 것이 검정의 프로세스입니다.

앞서 언급한대로, 이는 판사의 재판 프로세스와 유사한 면이 있어요. 근거를 모아서, 피고인의 유/무죄를 판단하는 것이죠. 실제 피고가 범죄를 저질렀는가에 해당하는 것이 모집단의 차이를 의미하고, 판결을 어떻게 하느냐가 귀무가설을 기각하느냐 마느냐에 해당하는 개념이라고 볼 수 있지요.

실제 / 판단	**범인 X**	**범인 O**
무죄	Good	오류
유죄	오류	Good

범인인 경우에 유죄 판결하거나, 범인이 아닌 경우 무죄 판결을 하는 것은 적절하게 잘 된 판결이라고 할 수 있겠으나, 실제 그렇지 않은 경우가 생길 수 있습니다. 잘못 판결하는 경우죠. 이 잘못된 판결에는 두가지 종류가 있어요. 범인을 무죄 판결하는 경우가 있겠고, 범인이 아님에도 유죄로 판결하여 징역을 살게 하는 경우가 있을 수 있어요. 둘 다 오류에 해당하는 개념이예요. 여러분은 어떤 오류가 더 중요하게 다루어져야 하는 오류라고 생각하나요? 여기서 중요하다는 것은 '더 조심해야 하는' 오류를 말합니다.

물론 두 가지 오류를 모두 줄이면 좋지요. 하지만, 동시에 두 오류 가능성을 줄이는 것은 불가능 하다는 것이 증명되어 있습니다. 하나를 줄이면 다른 오류가 늘어나게 되는 것이죠. 예를 들어 범인이 아닌 사람을 유죄판결하는 것이 그 반대의 경우보다 더 나쁜 일이라고 생각하여, 이에 대한 확률을 0으로 만들어야겠다고 해 봅시다. 모든 피고인에 대해 '무죄'로 판결한다면, '범인이 아닌 사람을 유죄로 판결하는 오류'는 생기지 않겠죠. 이 경우 '범인을 무죄로 판단하는 오류'가 증가하게 되죠. 반대로, 모든 피고인을 '유죄'로 판결한다면, '범인을 무죄로 판결하는 오류' 역시 발생하지 않지만, 무고한 사람을 유죄판결하는 오류는 증가하게 됩니다. 이러한 재판은 의미가 없겠죠.

법률을 다루는 판사에게도 '10명의 범인을 놓치는 한이 있더라도, 무고한 사람을 가두는 판결을 가장 경계해야 한다'라는 규칙이 있다고 합니다.

이러한 개념을 통계적 가설검정에서도 그대로 적용할 수 있어요.

실제 / 판단	**차이가 없음**	**차이가 있음**
차이가 없다고 판단	O (1-α)	X (type II error, false negatives, β)
차이가 있다고 판단	X (type I error, false positives, α)	O (statistical power, 1-β)

실제로(모집단에서) 차이가 없는데, 차이가 있다고 판단 (reject H0)하는 오류를 통계학에서는 type I error라고 합니다. False positive라고도 하고, 우리말로 하면 '제1종 오류'라고 해요. False positive라는 말은 '양성이라고 잘못 예측했다'라는 뜻으로 이해하면 됩니다. 반대로, 실제 차이가 있음에도 불구하고 차이가 없다고 판단(do not reject H0)하는 경우를 일컬어 type II error가 발생했다고 합니다. 이는 false negative에 해당하는 경우고, 우리말로는 '제2종 오류'라고 번역합니다. 앞서 재판의 예에서와 같이 어떤 오류가 더 중요하게 다루어져야 하는 오류일까요? 10명의 범인을 놓치더라고 한 명의 무고한 사람을 잡아 들이지 말라는 원칙으로 생각해 봅시다. 차이가 있는 10번의 검정을 '차이가 없다'고 판단할지라도, 모집단에서 차이가 없는 경우를 차이가 있다(유의하다)로 판단하는 것을 더 경계해야 한다는 뜻이 되겠지요. type I error가 더 조심스럽게 다루어야 하는 오류라는 뜻입니다.

통계적 검정의 목적은 '차이가 있는 것을 차이가 있다고 판단하는 것'을 목표로 하고 있지요. 그런데, 모집단에서 차이가 있는지 없는지는 분석단계에서 알 수 없는 일이죠. 모르기 때문에 가설검정을 하는 것이니까요. 따라서, 이 목적은 다음과 같이 수정되어 표현해야 합니다.

'가설 검정을 통해 차이가 있다고 판단한 것은 진짜로 모집단에서 차이가 있다'

가급적 위의 원칙이 지켜지도록 가설검정을 운영해야 한다는 뜻이예요.

이 말의 의미를 조금 생각해 봅시다. 통계학자가 약효가 있다고 판단한 약은 진짜로 효과가 있는 약이라는 말이예요. 이 말은 약효가 있다는 판결을 매우 보수적으로 한다, 웬만하면 약효가 있다고 판단하지 않는다, 내가 약효가 있다고 판단했다면 그 약은 진짜 효과가 있는 약이다. 이런 말이 되는 것이지요. 따라서, 통계적 검정은 매우 보수적인 판단입니다. 임상시험과 같은 과정에서 효과가 있다고 판단하면 이 약은 시판이 되어 국민들에게 보급됩니다. 만일 약효가 없는 약이 보급된다면, 이는 심각한 부작용을 초래합니다. 효과가 있는 약에 대해 효과가 없다고 판단하는 경우는 시판이 안되겠죠. 제약회사에 손해가 있을 겁니다. 하지만, 이는 감당할 수 있다고 보는 것이죠. 조금 더 근거를 강화해서 다시 허가 신청을 해도 되는 것이니까요.

때문에, 가급적 효과가 있다는 판단을 보수적으로 합니다. 일반적으로 type I error의 확률을 5%로 고정해 놓고 판단하는 이유죠. 이 type I error의 허용한계를 α라고 합니다. 앞서 신뢰구간 이야기할 때 나왔었죠? 모집단에서는 실제 차이가 없는 상황에서 이를 '유의하다', '차이가 있다'라고 판단하는 오류를 얼마나 허용할 것인가에 해당하는 문제죠. 이를 0으로 만들면, 모든 검정에서 '차이가 없다'라는 판단을 한다는 의미가 되니, 타당하지 않게 되는 것이죠. 때문이 이 확률을 적절한 수준에서 허용하여 분명 차이가 있는 것은 차이가 있다고 판단할 수 있도록 하자는 의미입니다.

유의수준을 5%로 둔다는 말은 type I error를 최대 5%까지만 허용한다는 뜻이예요.

자, 그렇다면 p-value의 의미를 생각해 봅시다. 이는 귀무가설하에서 양 극단에 위치할 확률을 계산한 것이죠. 이 확률이 작으면 귀무가설을 기각한다고 했습니다. 즉, **귀무가설하에서 현재의**

데이터와 같은 상황이 발생할 가능성이 작다는 의미라고 지난 시간에 강조했습니다. 따라서, reject H0를 결정하게 되는 것이죠.

이를 다른 방향으로 생각해 봅시다. 발생할 '가능성이 작다'라는 말은 '발생하지 않는다'는 아니죠. 발생할 수도 있다는 뜻입니다. 이러한 상황이 발생했다고 생각해 봅시다. 정말로 모집단의 평균이 160인데, 우연히도 1000명의 표본이 키가 큰사람 위주로 뽑힌 것이죠. 때문에 표본의 평균이 165가 나온 경우입니다. 표본의 평균이 180이라고 해 봅시다. 모집단의 평균이 160일 때 이러한 일은 잘 일어나지 않겠지만, 우연히 엄청 키가 큰 사람들이 표본으로 뽑히는 바람에 이런 상황이 발생할 수도 있는 것이죠. 그렇다면 reject H0라는 결정이 잘못된 것입니다. 하지만, 표본의 평균이 크면 클수록, 모집단의 평균이 160이라는 귀무가설을 유지하기는 어렵겠죠. 이를 확률화하여 계산한 것이 p-value라고 했습니다.

따라서, p-value라는 것은 귀무가설이 참일 때, 표본과 같은 상황이 발생할 확률입니다. 이는, 만일 귀무가설을 reject 했을 때, type I error가 발생할 확률과 같은 개념이죠. 귀무가설이 참일 때, 기각하는 것이 type I error라고 했으니, 귀무가설을 기각했을 때, 1종 오류가 얼마나 생길지를 확률로 표현한 것이 바로 p-value입니다.

1종오류의 허용한계를 '유의수준'이라 했습니다. 이를 일반적으로 5%로 둔다고 했죠. p-value는 귀무가설을 기각할 때, 발생할 1종 오류의 확률입니다. 따라서, p-value가 유의수준 5%를 넘는다면, 허용한계치를 넘는 것이므로 귀무가설을 기각하지 않고, 5%보다 작다면 귀무가설을 기각한다. 이런 원칙을 세울 수 있습니다.

결론적으로 검정과정을 정리하면,

1) 귀무가설과 대립가설을 세우고,
2) 적절한 분석법을 선택하여
3) 데이터를 활용해 검정통계량을 계산하고,
4) 계산된 검정통계량을 기초로 p-value를 계산한다.
5) p-value가 유의수준보다 작으면 귀무가설을 기각하고, 크면 기각하지 못한다.

이러한 과정으로 진행한다는 것입니다.

다음 강의에서 몇가지 예제를 풀어 봅시다.