알아두면 쓸모있는 통계관련 잡학상식

검증(檢證)과 검정(檢正)

검증하다와 검정하다는 한글로 쓰면 비슷해보이지만, 그 의미는 분명히 다르다. 게다가 검정하다는 검정(檢定)과 검정(檢正)의 두 가지로 사용된다. 이를 구분하기 위해 영어표현을 살펴보자.

검증(檢證)	verify	증명
검정(檢正)	test	실험
검정(檢定)	authorize	승인

우리가 하는 것은 사회조사를 통해 해당 사실을 test하는 것이다. 따라서 검정(檢正)하다가 정확한 표현이다. 이 test와 관련하여 칼 포퍼(Karl Popper)는 그의 반증이론(The theory of Falsification)을 통해 사회과학에서 100% 확실한 진리를 검증(檢證)하는 것은 불가능에 가까운데, 왜냐하면 언제든 설명되지 않는 부분에서 반증자료가 나타날 가능성이 있기 때문이라 말한다.
그래서 가설은 검정(test)하는 것이지, 검증(verify)할 수는 없다.

공(空)과 무(無), 0과 null

없다는 것을 나타내는 다양한 표현이 있다. 0, ○(공), 無 그리고 null 등 이들은 서로 어떻게 다른 것일까?
일단 우리는 숫자 0을 영(零)과 공(空)으로 읽는다. 하지만 원칙은 '영'으로 읽는 것이 맞다. 굳이 구분하자면 공은 ○과 같은 기호로 보는 것이 옳을 듯하다.
하지만 이것은 무(無)와는 조금 다르다. 불교에서는 색즉시공(色卽是空)이라 말한다. 있는 것(色)이 어떻게 없는 것(空)이 될 수 있을까라는 오묘한 철학적 논쟁은 잠시 뒤로 미뤄두고 그 표현만 가져와보자. 있는 것이 없어졌다면 그것은 없는 것(無)인가 없어진 것(空)인가? 당연히 후자일 것이다. 애초에 없는 것을 무(無)라 하고, 없어진 것을 공(空)이라 한다.
그런데 이런 개념은 신기하게도 프로그래밍에서도 등장한다. 바로 null과 0이다. 예를 들어 종이에 0이라는 숫자 하나를 썼다고 가정해보자. 이는 숫자 0이 있는 것이다. 반면 아무 것도 쓰여지지 않은 빈 종이를 null이라 할 수 있다.
즉, 무(無)는 null에, 영(零) 또는 공(空)은 0에 대입할 수 있을 것이다. 그렇다면 null hypothesis는 표현 그대로라면 귀무(歸無)가설이라고 말할 수 있겠지만, 차이가 없다는 말이 null이라는 뜻은 아니니 영(零)가설이 더 타당하지 않나 싶기도 하다.

부등식의 표현 이해

우리는 초등학교 때 부등식에 대해 배웠다. 그리고 나이가 들면서 미만/이하, 초과/이상의 구분은 기억하고 있다. 하지만 오히려 이를 우리말로 표현하면 헷갈려한다.

p<.05	미만	p는 0.05보다 작다.
p≤.05	이하	p는 0.05보다 작거나 같다.
		p는 0.05보다 크지 않다.

통계에서 영가설의 기각 여부를 판단하는 기준으로 유의확률 p<.05와 같이 표현하곤 한다. 이는 p value가 0.05보다 작다는 말이지만, 0.05보다 크지 않다는 뜻은 아니다.

덧붙여 부등호 중 '작거나 같다' 또는 '크거나 같다'의 표기는 ≤와 ≥를 사용한다. 하지만 예전에 수학을 배우신 분들은 ≦와 ≧가 더 익숙할 것이다. ≤과 ≦, ≥과 ≧은 같은 의미이다. 그리고 오늘날은 ≤과 ≥를 사용한다.

p value(유의확률)를 표기하는 방법

2010년의 APA((American Psychological Association) style manual 제6판(p.141)에 따르면, *p* value를 다음과 같이 표기하라고 말한다.

① 소숫점 앞의 0은 표기하지 않는다. 예) 0.051(X) .051(O)
② 소숫점 셋째자리까지 직접 기술한다. 예) p=.051
③ 만일 .000보다 더 작다면(예를 들어 p=.000123), p<.001로 표기한다.
※ SPSS의 경우 버전 26까지는 .000으로, 버전 27부터는 p<.001로 나타낸다.

덧붙여 몇 가지 주의사항을 언급해보자면,
④ 통계에서 쓰는 기호는 기본적으로 이탤릭체로 쓰며, 사이띄우기는 하지 않는다.
⑤ 또한, "유의미하다(significant)"의 반대말은 "무의미하다(insignificant)"가 아니라, "유의미하지 않다(nonsignificant)"이다.

p value(유의확률)와 통계량

p value와 통계량은 연결되어 있다. 어쩌면 당연한 이야기이다.
하지만 직접 계산하여 통계량을 산출해 본 경험이 없이 통계 프로그램이 보여주는 결과값만을 확인해본 것이 전부라면, 따로따로 제시되는 유의확률과 통계량을 별도의 것으로 이해하는 경우도 생길 수 있다.
t test를 통해 통계량 t=2.127로 나와다면 무조건 *p*<.05일 수밖에 없다. 만일 당신이 통계적 유의미성만 확인하고자 한다면, 통계량 또는 *p* value 둘 중 하나만으로도 충분히 그 결과를 해석할 수 있다.

왜 유의확률(p value)은 0.05를 기준으로 하는가?

이 0.05라는 값은 통계적 유의미성을 지지하는 기준값이다. 즉 p<.05이면, 영가설 기각이 통계적으로 유의미하다는 뜻이 된다. 하지만 의문이 들지 않는가? 왜 하필 0.05일까? 만일 내가 한 연구에서 유의확률이 0.051이 나왔다면 좀 아깝지 않을까?[1)]

사실 이 0.05라는 값은 반드시 0.05이어야 할 과학적 근거가 있는 것은 아니다. 다만, 20세기 위대한 통계학자 중의 한명인 Ronald Fisher가 1925년 그의 저서 『Statistical Methods for Research Workers(p.46)』에서 처음 언급하게 된다.[2)]

> "The value for which P = 0.05, or 1 in 20, is 1.96 or nearly 2 ; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not."
>
> "P = 0.05, 즉 20분의 1인 값은 1.96(대략 2)입니다. 편차가 중요한지 여부를 판단할 때 이 점을 한계로 삼는 것이 편리합니다."

귀납법이 갖는 철학적 한계를 해결하기 위해 통계적 접근방법을 활용한 것으로, 현대 통계의 역사를 다룬 『The Lady Tasting Tea: How Statistics Revolutionized Science in the 20th Century』에서 David Salsburg(2001)는 Fisher의 결정이 '임의적' 인 것이었다고 말한다.

p<.05는 관행일 뿐 절대시할 수치는 아니라 할 것이지만, 사회적 약속인 것 또한 사실이다.

---

1) p-hacking, p 해킹에 대해 더 찾아 읽어보자.
2) 위키피디아 https://en.wikipedia.org/wiki/Ronald_Fisher

확률(probability)과 가능도(likelihood)

동전이나 주사위를 던졌을 때 나올 수 있는 경우의 수는 각각 동전 1/2, 주사위 1/6 이 되며, 각 경우의 수의 합은 1이 된다. 동전의 경우 앞면이 나올 확률은 1/2이라 알고 있다. 이처럼 경우의 수가 정해져 있는 분포를 '이항분포'라 하고 이런 확률의 분포를 '이산확률분포'라 부른다.
여기까지는 매우 직관적이다. 하지만 좀더 엄밀히 살펴보려면, 두 가지를 확인할 필요가 있다.

- **실제로 동전의 앞면이 나올 확률이 1/2인가?**
- **이산확률분포가 아닌 경우는 어떠한가?**

실제로 동전을 던졌을 때 10번을 던져도 앞면이 한번도 나오지 않는 경우도 있을 수 있다. 하지만 100번을 던진다면 어떨까? 앞면이 나올 확률은 1/2로 수렴할 것이다. 그래서 우리는 이 1/2을 기댓값이라 부른다.

평균(mean) = 기댓값(E(X))

한편 연속인 수라면 어떨까? 0과 1 사이에도 무수히 많은 숫자들이 있다. 따라서 특정 사건이 일어날 확률은 1/∞이 되어, 항상 0에 수렴한다. 이로 인해 확률을 구하는 것이 의미가 없다. 때문에 특정 구간에 속할 확률(확률밀도함수, Probability Density Function, PDF)을 사용한다.
예를 들어 1에서 5 사이의 숫자가 뽑힐 확률은 알 수 없지만, 1에서 2사이의 숫자가 뽑힐 확률은 1/4, 즉 25%가 된다. 그리고 이를 비교하는 방법이 가능도(likelyhood)이다.
셀 수 있는 사건 즉 이항분포라면 가능도와 확률은 같다. 앞서 예에서 1, 2, 3, 4, 5의 숫자가 선택지인 경우 특정 숫자 2가 나올 확률은 1/5, 즉 0.2가 된다. 하지만 연속인 사건이라면 확률이 0이기 때문에 가능도는 확률밀도함수(PDF) 값이 된다. 마찬가지로 1~5 사이의 선택지는 무한하지만 1~2사이의 숫자가 뽑힐 가능도는 1/4, 즉 0.25가 된다.
그리고 그 가능성이 최대가 되는 값이 최대가능도 추정량(Maximum Likelihood Estimator, MLE)이, 위 예에서 MLE=0.25이다. 단, 이는 정규분포를 전제한다.

Z 점수(Z score, 표준점수)

원점수를 표준편차 단위로 변환한 것으로 평균에서 얼마나 떨어져 있는지를 나타내준다. 또한 2개의 다른 표본 혹은 모집단으로부터 측정된 변수값은 이들을 직접 비교할 수 없는데, Z 점수(표준점수)로 변환하면 이들 간의 비교가 가능해진다.

$$z = \frac{x-\mu}{\sigma} \quad \text{또는} \quad Z = \frac{X-\overline{X}}{s}$$

평균 추론에 필요한 조건

통계에 대해 공부하다보면, 헷갈리는 것 중의 하나가 모집단에 사용되는 기호와 표본집단에 사용되는 기호가 혼재해서 사용된다는 점이다. 바로 평균과 표준편차에 대한 것이 그것이다.

모집단		표본집단	
모평균	μ	표본 평균	$\overline{x}$
모표준편차	σ	표본표준편차	s

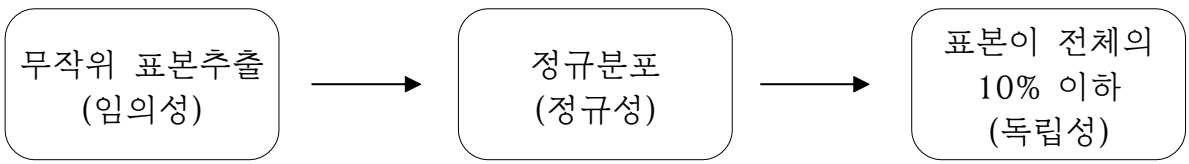
일반적으로 모집단의 평균과 표준편차보다는 표본집단의 평균과 표준편차에 대해 확인하는 것이 훨씬 쉽다. 그리고 당연히 이 표본이 모집단을 대표할 수 있다는 확신을 전제한다.[3)]

평균 추론에 대한 필요조건을 충족했을 때, 우리는 표본이 모집단을 대표할 수 있다고 보고 이때 모표준편차 σ는 표본표준편차 s로 대체할 수 있다. 그리고 그 조건은 다음과 같다.

첫째, 임의성이다. 표본은 무작위로 추출되어야 한다.

둘째, 일반성이다. 표본분포는 정규분포를 따라야 한다. 왜도의 절댓값이 2보다 작고, 첨도의 절댓값이 7보다 작을 때 정규성을 가정한다.

셋째, 독립성이다. 각각의 관측값은 독립이어야 한다. 표본의 수는 모집단의 수의 10% 이하로 관측값을 제거해도 모집단에 영향을 미치지 않아야 한다.



3) 물론 아닌 경우도 분명 있다.

표본이 정규분포(정규성)인지 여부를 어떻게 알 수 있는가?

표본의 정규분포를 확인하는 방법은 다양하다. 히스토그램과 같은 도표를 보고 추정할 수도 있고, 왜도와 첨도, 또는 정규성 검정을 통해 확인할 수도 있다.
정규성 검정은 표본의 크기에 따라 두 가지로 나뉜다.
첫째, $n \geq 50$이면, Kolmogorov-Smirnov(콜모고로프-스미르노프) 검정(ks test)을 통해 확인한다.
둘째, $n<50$이면, Shapiro-Wilk(샤피로-윌크) 검정을 통해 확인할 수 있다.
검정 결과 $p>.05$이면, 정규성을 가정한다. 다만 이 두 검정은 매우 엄밀한 검정으로 정규성을 가정하는 경우를 확보하는 것이 쉽지 않다.
이에 조금 더 유연한 방법이 왜도와 첨도를 확인하는 방법이다. 기술통계를 통해 확인되는 왜도(skewness)의 절댓값이 2보다 작고, 첨도(kutosis)의 절댓값이 7보다 작으면 정규성을 가정한다.

	왜도의 절대값	첨도의 절대값
West et al(1995)[4]	2.0 미만	7.0 미만
Hong et al(2003)[5]	2 미만	4 미만
Kline(2005)[6]	3.0 미만	8.0 미만

데이터분석을 실시하기에 앞서 표본의 정규성을 먼저 확인하는 것이 전제되어야 모수통계 검정을 실시할지 비모수통계 검정을 실시할지를 결정하게 된다.
한편 중심극한정리(central limit theorem, CLT)를 예로 들어 표본의 수(n)가 30 이상이면 정규성을 가정한다는 주장이 있는데, 이는 틀린 표현이다. 중심극한정리는 '표본집단' 의 수가 증가함에 따라 모집단의 분포가 정규분포에 근접해 가는 현상을 말한다. 즉 표본집단의 수가 판단의 기준이다. 표본의 수를 말하는 것이 아닌데, 이를 오해해 잘못 전하는 경우가 있어 주의가 필요하다.

$n \geq 30$이라고, 정규성을 가정하는 것은 아니다.

4) West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues, and applications (p. 56–75). Sage Publications, Inc.
5) Hong, S., Malik, M. L., & Lee, M.-K. (2003). Testing configural, metric, scalar, and latent mean invariance across genders in sociotropy and autonomy using a non-Western sample. Educational and Psychological Measurement, 63(4), 636–654.
6) Kline, T. J. (2005). Psychological testing: A practical approach to design and evaluation. Sage Publications.

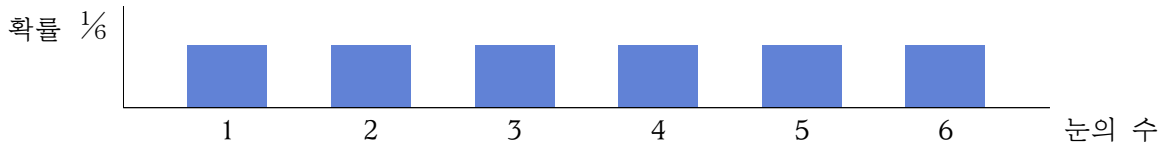
중심극한정리(central limit theorem, CLT)

중심극한정리는 동일한 확률분포를 가진 독립 확률 변수 n개의 평균의 분포는 n이 적당히 크다면 정규분포에 가까워진다는 정리이다. 그리고 이를 간단히 표현해 보자면, "표본의 평균들의 분포는 정규분포를 따른다" 이다.

모집단		표본집단$_1$		표본집단$_2$		표본집단$_m$	
모평균	μ	표본 평균	$\overline{x_1}$	표본 평균	$\overline{x_2}$	표본 평균	$\overline{x_m}$
모표준편차	σ	표본 표준편차	s_1	표본 표준편차	s_2	표본 표준편차	s_m

앞서 표본의 수(n)가 30 이상이면 정규성을 가정한다는 주장이 틀린 표현이라 언급한 바 있다. n≥30이 아니라, 표본집단$_m$≥30이면 이 표본집단의 평균들($\overline{x_1}$, $\overline{x_2}$, … , $\overline{x_m}$)의 분포가 정규분포를 따른다는 말이다.

상식적으로 주사위를 굴렸을 때 나올 수 있는 눈의 경우의 수는 1/6로 정규분포를 따르지 않는다.

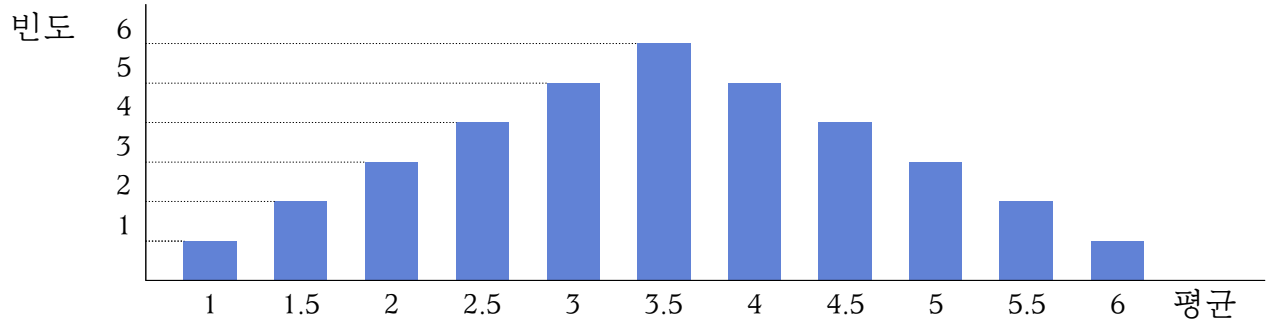


그리고 여기서 뽑은 표본의 분포 또한 정규분포를 따르지 않는다. 하지만 표본의 수(n)이 커지면, 모집단의 분포와 닮아갈 것이다. 실제로 주사위를 10번 굴렸을 때의 분포보다 100번 굴렸을 때의 분포가 더 모집단의 분포를 닮는다. 물론 이때의 그래프 또한 정규분포는 아니다.

그런데 만일 주사위를 두 번씩 굴린다면 어떻게 될까? 이때의 경우의 수는 (1,1), (1,2), … (1,6), (2,1), (2,2) … (2,6), … (6,6)까지 총 36개의 경우의 수가 나올 수 있다.

평균	1	1.5	2	2.5	3	**3.5**	4	4.5	5	5.5	6
빈도	1	2	3	4	5	**6**	5	4	3	2	1

그 평균과 빈도를 정리해 그래프를 그려보면 아래와 같이 나타나는데, 정규분포와 비슷한 모양을 나타내는 것을 볼 수 있다.



이렇게 주사위를 굴리는 횟수를 증가시키면 평균과 빈도의 그래프는 더욱 정규분포를 그리게 될 것이다. 중심극한정리는 이처럼 주사위를 30번 이상 굴리게 되면, 정규분포에 수렴한다는 사실에 대한 증명이다.

정리해보자면, 표본 n의 크기가 증가하면, 표본의 분포는 모집단의 분포와 가까워진다. 한편 표본집단의 개수 m이 증가하면, 표본집단의 평균에 대한 분포는 정규분포에 가까워진다.

> 표본의 수가 증가하면 표본의 분포는 모집단의 분포에 가까워진다.
> 표본집단의 개수가 증가하면, 표본집단의 평균 분포는 정규분포에 가까워진다.

부트스트랩(Bootstrap)

앞서 표본의 정규성에서 설명한 중심극한정리는 표본집단을 30번 이상 추출한 경우 정규성을 가정한다는 말이다. 하지만 현실적으로 표본을 30번 이상 뽑아서 연구를 진행한다는 것은 불가능에 가깝다. 이에 등장한 개념이 부트스트랩(bootstrap)[7]이다. 부트스트랩이란 연구자가 추출한 표본집단을 모집단으로 여기고 이 표본에서 '복원추출' 의 방식으로 계속 반복해서 표본을 재추출하는 방법을 말한다.

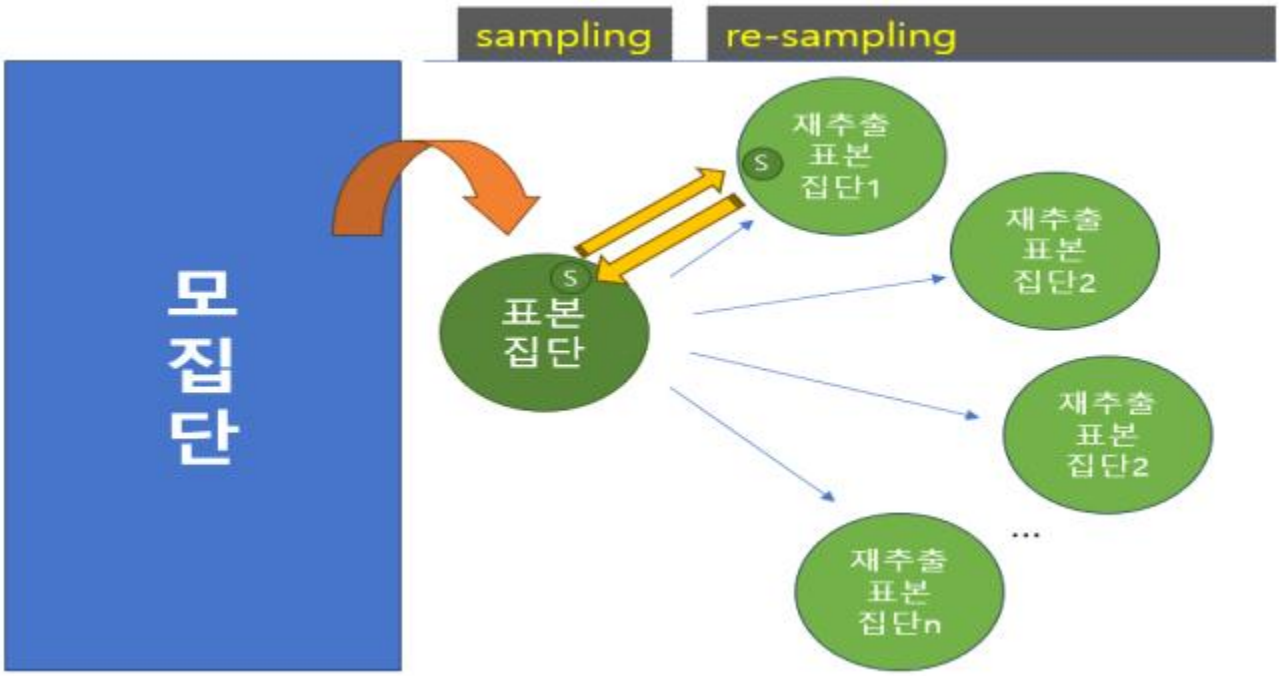


그림 2 복원추출 방법

복원추출이란 표본집단에서 표본 1개를 추출한 후 이를 다시 표본집단에 넣은 후 반복적으로 표본을 추출해 표본집단의 크기가 재추출표본집단의 크기와 같아질 때까지 시도하는 방법이다.

이때 재추출(re-sampling)의 횟수는 5,000회 또는 10,000회를 권장하고 있다.

7) 부트스트랩(bootstrap)은 옛날 장화(부츠, boots)에는 지퍼가 없어 통으로 신어야 해 매우 불편했다. 그런데 누군가가 장화 뒤에 작은 끝을 달아 잡고 신기 편하도록 만들었는데, 이를 bootstrap(장화끈)이라 불렀다., 이처럼 사소하지만 굉장히 실용적인 아이디어 또는 일사천리로 진행되는 과정을 가리키는 표현으로 사용된다. 통계에도 중심극한정리를 대체할 수 있는 좋은 아이디어라는 의미로 이런 표현을 차용해서 쓰고 있다.

독립변수와 종속변수: 인과관계와 변수

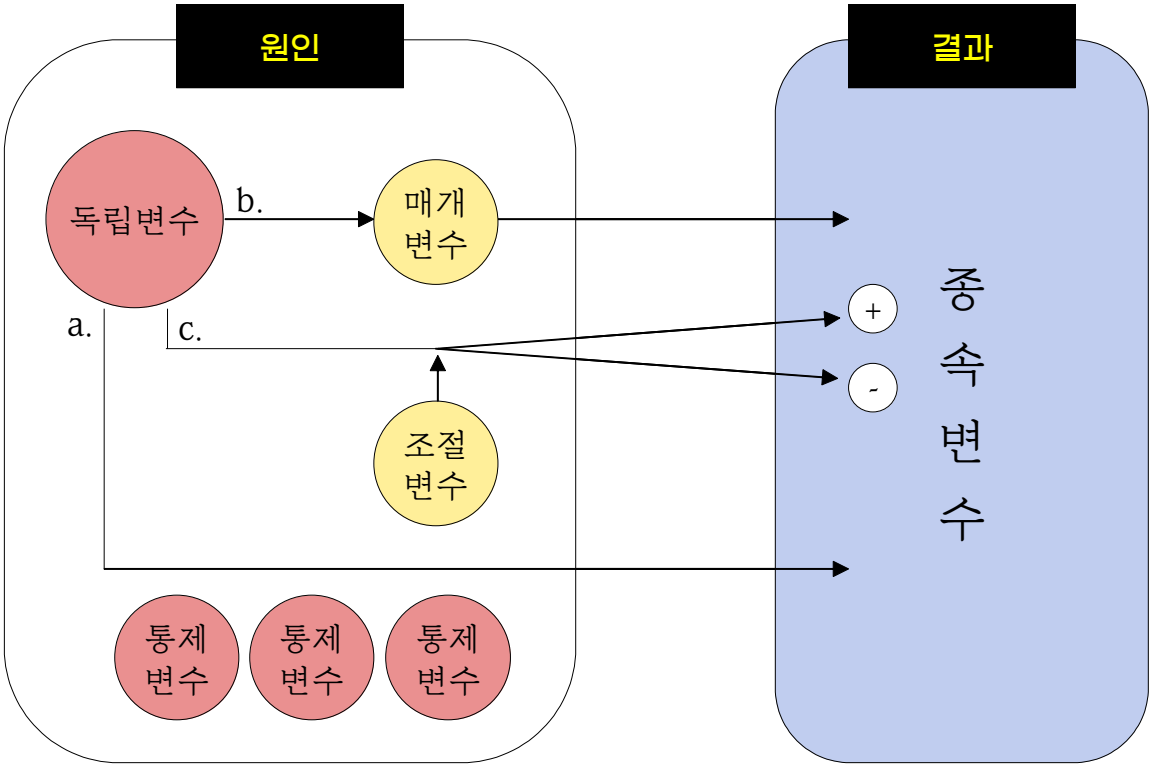


그림 3 원인과 결과의 관계를 밝히는 세 가지 방법

가장 단순히 설명하자면 독립변수는 원인이 되는 변수, 종속변수는 그로 인해 변하게 되는 결과변수라고 말할 수 있다(a). 하지만 이런 인과관계에 있어 단일 변수 하나가 절대적인 영향력을 미친다고 사회학에서 말할 수 있을까? 우리 사회에서 보여지는 수많은 현상들은 그만큼 많은 원인들이 얽히고 설켜 만들어낸 결과물일 것이다.
독립변수가 종속변수의 직접적인 원인이라기 보다는 다른 변수를 통해 영향을 미친다면 우리는 이를 매개효과라 하고 이런 변수를 매개변수라 부른다(b). 한편 독립변수가 원인이 맞지만 또다른 변수를 거칠 때 그 효과가 더 커지기도 하고 줄어들기도 하는데, 이를 조절효과라하고 그 변수를 조절변수라 한다(c).

한편 우리가 신이 아닌 다음에서야 모든 원인이 되는 변수들을 밝힌다는 것은 어려운 일이 아닐 수 없다. 나아가 종속변수에 영향을 미치는지 여부를 모두 확인한다는 것은 불가능하다. 하여 연구자는 본인이 변수로 내세운 것에 대해서만 얘기할 수 있다. 이때 만일 연구자가 특정 변수의 개입 여부와 상관없이 동일한 결과가 나타난다고 얘기하고 싶다면 어떻게 해야할까? 그 변수도 독립변수로 넣어서 분석하고 이와 상관없이 동일한 결과가 나타났다고 말하면 될 것이다. 이를 '통제변수'라 부른다. 예를 들어 키가 취업에 미치는 영향을 조사했다고 해보자.[8] 그리고 키가 크면 취업

8) 예시 참조 https://m.blog.naver.com/mj0147won/221155461247

이 더 잘된다는 가설을 세울 수 있을 것이다. 하지만 이렇게 생각해볼 수도 있을 것이다. 남자가 여자보다 평균키가 더 크기 때문에 이런 결론이 나온 것은 아닐까? 그렇다면 우리는 남자 따로 여자 따로 조사하고 그럼에도 불구하고 같은 결과가 도출되었다고 말함으로써 성별이라는 변수를 통제할 수 있다.

한편 통제변수가 독립변수가 종속변수에 미치는 영향을 통제하는 것이 틀린 말은 아니지만, 보다 엄밀히 말하자면 선행변수가 후행변수에 미치는 영향을 통제하는 것이다.[9] 때문에 이런 통제변수는 다분히 보조적인 것으로 연구의 핵심은 아니다. 따라서 굳이 통제변수가 미치는 영향을 해석해 논문이나 보고서에 포함할 필요 또한 없으며, 이는 오히려 연구의 논점을 흐리게할 수도 있어 지양해야한다.

그리고 일반적 특성을 통제변수로 사용해 독립변수에 미치는 영향을 통제한다고 말하는 것도 틀린 말이다. 만일 이같은 통제변수의 개념을 명확히 이해하지 못했다면 굳이 연구에서 이를 구분하지 말아야 할 것이다. 그냥 다양한 독립변수를 검토한 것일 뿐이다.

제3의 변수를 통제하는 것이 연구의 타당도를 높일 때만
통제변수를 사용하자.

9) 박원우·고동운·윤은성. 2010. "연구의 인과성 제고: 통제변수의 의의, 활용현상 분석 및 제언", 『노사관계연구』, 21, pp.1-49

측도와 척도

일반적으로 변수(variable)의 속성으로 구분한 기준을 말하는 것은 척도(scale)이다. 한편 변수를 측정한 값은 측도(measure)라 한다.

	측도(measure)	척도(scale)
정의	변수를 측정한 값(정보)의 특성	변수의 특성 및 이를 측정하는 기준
예시	평균, 최빈값, 중앙값/중위수 등	명목, 서열, 등간, 비율

리커트(Likert) 척도는 서열척도인가?

사회조사 논문을 보면 리커트 척도를 등간척도로 다루어 상관분석과 회귀분석을 해놓은 것을 많이 볼 수 있다. 하지만 뭔가 이상하지 않은가? 분명 배운대로라면 리커트 척도는 서열척도가 맞다. 이에 대한 논쟁은 과거부터 꾸준히 있어왔던 듯하다. 이에 대해 잘 정리한 논문이 있어 한편 소개한다.

바로 후이핑 우(Huiping Wu)와 싱온 렁(Shing-On Leung)이 2017년 Journal of Social Service Research에 기고한 "Can Likert Scales be Treated as Interval Scales? - A Simulation Study"이다. 이 연구의 Introduction을 보면, Jamieson(제이미슨, 2004)을 비롯한 인용해 엄밀히 말해 서열척도인건 분명하다고 본다. 한편 척도를 만들어낸 Stevens(스티븐스, 1946) 또한 서열척도를 등간척도로 다루었을 때의 유용성에 동의했다며, 리커트 척도의 개수를 늘려간다면 연속적인 척도로 보아 산술연산을 하는 것도 가능하다는 입장도 소개한다. 또한 Borgatta(보가타) & Bohrnstedt(보른스테드)는 리커트 척도를 불완전한 등간척도라 부르기도 한다.

서열척도를 등간척도로 다루는 것은 기본 가정을 위반한다는 문제점에도 불구하고 그 실효성이 높다는 딜레마를 안고 있다. 이 논문의 저자들은 그렇다면 얼마나 리커트 척도를 늘려가야 등간척도와 유사한 결과를 얻을 수 있는지에 대해 실험하고 그 결과 0~10까지 11점 척도가 된다면 등간척도로 보아도 무방한 결과를 도출한다고 결론내리고 있다.

무작위(random) 표본추출(표집, sampling)

무작위 또는 랜덤(random)로 표본을 뽑았다고 하면, 정해진 방식이 없이 표본이 우연히 선택되는 것을 상상하곤 한다. 이는 무작위 표본 추출(표집)에 대한 대표적인 오해이다. 이처럼 우연히 선택되는 것을 우리는 편의표집(convenience sampling)이라 한다. 기본적으로 표집 방법은 확률표집과 비확률표집으로 구분되며, 비확률표집에서도 최후의 보루가 바로 편의표집이다.
바람직한 표집은 표본이 편향되지 않게(unbiased) 선택될 것을 전제로 한다. 그리고 이를 위한 가장 대표적인 방법이 무작위(random)이다.
그렇다면 무작위 표본추출은 어떤 식으로 이루어지는가? 대표적인 방법 중의 하나인 난수표를 사용하는 방법은 아래와 같은 절차로 진행된다.

① 모집단에 모든 개체에 같은 자리수의 일련번호를 부여한다.
② 난수표에서 시작점을 무작위로 정하여 그 시작점으로부터 개체에 부여된 번호와 같은 자릿수만큼의 숫자를 읽어나간다.
③ 반복해서, 같은 자리수의 숫자를 고르되 반복된 숫자나 일련번호로 주어진 숫자 외에는 제외한다. 이 때, 원하는 표본의 크기만큼 숫자를 고른다.
④ 뽑힌 숫자의 일련번호를 가진 개체를 표본으로 정한다.

무작위 표본추출이란 모집단의 각각의 요소 또는 사례들이 표본으로 선택될 가능성이 같게 되는 표본 추출법이다. 이론적으로는 가장 단순한 표본추출법이지만 선택 가능성을 같게 하기 위해서는 모든 개체가 추출 이전에 확인되어야 하고 표시되어야 하기 때문에 비용이 많이 들고 실현 가능성이 적다는 문제점이 있다.

표본오차(sampling error)

일반적으로 우리는 모집단이 아닌 표본집단을 대상으로 분석을 실시한다. 하지만 이런 표본집단과 모집단 사이에는 작은 차이가 존재할 수밖에 없다. 이런 차이를 표본오차(sampling error)라 부른다.

예를들어 지난 20대 대통령선거 지상파 3사의 출구조사 결과를 살펴보자.

신뢰수준 95% 표본오차 ±0.8%p[10]	윤석열 후보 48.4% 이재명 후보 47.8%

이를 하나하나 해석해보면 다음과 같다.

첫째, 신뢰수준 95%는 같은 조사를 100번 했을 때 95번은 같은 결과가 나올 것이라 기대할 수 있다는 뜻이다.

둘째, 표본오차 ±0.8%p는 윤석열 후보의 실제 득표율이 47.6%~49.2%, 이재명 후보의 득표율은 47.0%~48.6% 사이에서 결정될 것으로 기대된다는 의미이다.

그리고 이 말은 출구조사의 결과만 놓고 살펴본다면, 누가 최종적으로 대통령이 될지에 대한 예측은 되지만 결과값이 오차범위 내에 있다는 뜻이다.

한편 이런 표본오차(e)는 표본의 크기와 관련이 있다.

e: 표본오차(sampling error) N: 모집단의 크기(Population Size) n: 표본의 크기(Sample Size) Z: 95% 신뢰수준이면 Z=1.96, 99% 신뢰수준이면 Z=2.54 P: 관찰치(the observed percentage), 모집단을 모르는 경우 최대 표본오차를 구하기 위해서 P=0.5를 사용한다.	$e = Z \times \sqrt{\frac{P(1-P)}{n}}$ $e = 1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} = 1.96 \times \sqrt{\frac{1}{4n}}$

그리고 이런 표본오차는 오차한계(margin of error), 최대허용오차, 오차범위, 표집오차 등과 같은 의미로 사용된다.

표본오차 = 오차한계 = 최대허용오차 = 오차범위 = 표집오차
↳ margin of error

10) %와 %p
- 퍼센트(%): 백분율, n/N*100
- 퍼센트포인트(%p): 퍼센트 간의 차이

예) 제19대 대선 투표율은 77.2%였고, 제20대 대선 투표율은 77.06%였다. 투표율은 지난 대선 대비 0.14%p 감소한 것으로 나타났다.

제1종 오류와 제2종 오류[11)]

재1종 오류와 제2종 오류의 개념을 정립한 것은 통계학자 예르지 네이만과 이건 피어슨이다.

1928년, 오류의 두 가지 원인을 "채택해야할 가설을 기각하는 오류" 와 "기각해야할 가설을 채택하는 오류" 로 정의[12)]하였고, 1930년, "가설 검증에서는 다 두 가지를 늘 생각해야만 한다. (1) 우리는 참인 가설을 기각해버릴 가능성을 최대한 낮게 해야만 한다. (2) 거짓으로 생각되는 가설이 기각되도록 검증해야만 한다." 고 말하였다.[13)]

… in testing hypotheses two considerations must be kept in view, (1) we must be able to reduce the chance of rejecting a true hypothesis to as low a value as desired; (2) the test must be so devised that it will reject the hypothesis tested when it is likely to be false.

그리고 1933년, 그들은 이런 오류를 두 종류로 나누고, 각각 제1종 오류와 제2종 오류라 이름붙였다.[14)]

…these errors will be of two kinds:
(I) we reject H0 when it is true,
(II) we accept H0 when some alternative hypothesis HA or H1 is true

이러한 오류들은 다음의 두 종류로 나뉜다.
(I) 귀무가설 H_0가 참인데도 기각하는 경우
(II) 대립가설 H_A또는 H_1이 참인데도 H_0를 채택하는 경우

11) 출처: 나무위키

12) Neyman, J. & Pearson, E.S., "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference, Part I", reprinted at pp.1−66 in Neyman, J. & Pearson, E.S., Joint Statistical Papers, Cambridge University Press, (Cambridge), 1967 (originally published in 1928), p.31)

13) Pearson, E.S. & Neyman, J., "On the Problem of Two Samples", reprinted at pp.99−115 in Neyman, J. & Pearson, E.S., Joint Statistical Papers, Cambridge University Press, (Cambridge), 1967 (originally published in 1930), p.100

14) Neyman, J. & Pearson, E.S., "The testing of statistical hypotheses in relation to probabilities a priori", reprinted at pp.186−202 in Neyman, J. & Pearson, E.S., Joint Statistical Papers, Cambridge University Press, (Cambridge), 1967 (originally published in 1933)

통계에서 ‘로버스트(robust)’의 의미

robust를 사전에서 찾아보면, 왕성한, 팔팔한, 튼튼한, 강력한, 건장한 등의 뜻으로 해석된다.
하지만 통계에서는 이 robust를 ‘더욱 적절한’의 뜻으로 사용한다. 예를 들어 집단의 특징을 대표할 때 일반적으로 평균을 사용하지만 경우에 따라서는 중위수/중앙값이 더 나을 때가 있다. 우리나라 국민의 소득수준이 그렇다. 평균소득과 중위소득 중 중산층을 더 잘 대표하는 것은 중위소득이다. 이때 중위수는 평균보다 더 robust하다.
또한 통계를 하다보면 이상값(outlier)가 존재하는 경우 데이터가 왜곡되는데, 이러한 ‘이상값/극단값에 영향을 덜 받는’ 통계량을 robust하다고 말한다. 예를 들어 일원배치분산분석(ANOVA)을 할 때, 분산의 동질성 검정 결과 데이터의 등분산이 가정되지 않는다면, 평균의 동질성에 대해 robust한 검정을 사용해야한다. 그런 방법이 Brown-Forsythe ANOVA 또는 Welch ANOVA이다.

Mann Whitney U = .000[15)]

Mann Whitney U 검정통계량이 .000인 경우가 있다. 이건 도대체 무슨 의미일까?
Mann Whitney U 검정은 순위검정이다. 즉 각 값들을 등수화(서열화)하고 일련의 수식을 통해 각 집단의 U 통계량을 구한 후, 그 차이를 또다른 식을 통해 검정하는 방식으로 계산한다.
U_max 통계량은 두 집단에서 계산된 U값 즉 U_1, U_2 중 더 큰 값으로 한다. 그리고 이를 표준화하는 과정에서, 이(U_max)에서 U의 평균값(m=(U_1+U_2)/2)을 뺀 값을 다시 표준편차로 나눈값(z)으로 한다. 이때 표준편차가 지나치게 커지거나, U_max와 m의 차이가 없으면 해당값은 0이 된다. 예를 들어 한 집단의 데이터가 다른 집단의 데이터보다 모든 경우에 월등히 작을 때가 그렇다.
하여튼 해당값이 0이 나왔다고 의미가 없다거나 틀렸다는 뜻은 아니다. 단지 유의확률을 유의수준과 비교해 그대로 해석하면 된다. z값은 표준화를 위한 통계량으로 보정을 위해 사용되는데, 일반적으로 크게 영향을 미치지 않아 특별한 경우가 아니라면 신경쓰지 않아도 된다.
Wilcoxon의 W는 Mann-Whitney-Wilcoxon 검정으로 Mann Whitney U 검정과 유사한, 즉 학자가 한명 더 붙어서 수식을 확장시켰다고 보면 된다.

15) 출처: https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test